Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciências
Instituto de Matemática e Estatística

Jorge Luiz de Jesus Goulart

Avaliação de Usabilidade e Aprendizagem de Algoritmos com o uso do TuPy Online

Rio de Janeiro
2019

Jorge Luiz de Jesus Goulart

Avaliação de Usabilidade e Aprendizagem de Algoritmos com o uso do TuPy Online

Dissertação apresentada como requisito para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciências Computacionais, da Universidade do Estado do Rio de Janeiro.

Orientador: Prof. Dr. Fabiano de Souza Oliveira

Orientador: Prof. Dr. Paulo Eustáquio Duarte Pinto

Rio de Janeiro

2019

CATALOGAÇÃO NA FONTE

UERJ/REDE SIRIUS/BIBLIOTECA CTC/A

G694 Goulart, Jorge Luiz de Jesus.
Avaliação de usabilidade e aprendizagem de algoritmos com o uso de TuPy online/ Jorge Luiz de Jesus Goulart. – 2019.
88 f.: il.

Orientadores: Fabiano de Souza Oliveira, Paulo Eustáquio Duarte Pinto
Dissertação (Mestrado em Ciências Computacionais) - Universidade do Estado do Rio de Janeiro, Instituto de Matemática e Estatística.

1. Algoritmos - Teses. 2. Estatística não paramétrica - Teses. I. Oliveira, Fabiano de Souza. II. Pinto, Paulo Eustáquio Duarte. III. Universidade do Estado do Rio de Janeiro. Instituto de Matemática e Estatística. IV.Título.

CDU 510.5

Patricia Bello Meijinhos CRB7/5217 - Bibliotecária responsável pela elaboração da ficha catalográfica

_____ _____

Assinatura Data

Jorge Luiz de Jesus Goulart

Avaliação de Usabilidade e Aprendizagem de Algoritmos com o uso do TuPy Online

Dissertação apresentada como requisito para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciências Computacionais, da Universidade do Estado do Rio de Janeiro.

Aprovada em 16 de dezembro de 2019.

Banca Examinadora:

Prof. Dr. Fabiano de Souza Oliveira (Orientador)
Instituto de Matemática e Estatística - UERJ

Prof. Dr. Paulo Eustáquio Duarte Pinto (Orientador)
Instituto de Matemática e Estatística - UERJ

Prof.ª Dra. Vera Maria Benjamim Werneck
Instituto de Matemática e Estatística - UERJ

Prof. Dr. Igor Machado Coelho
Instituto de Matemática e Estatística - UERJ

Prof. Dr. Antonio Orestes de Salvo Castro
Instituto de Matemática e Estatística - UERJ

Prof. Dr. Danilo Artigas da Rocha
Universidade Federal Fluminense

Rio de Janeiro
2019

AGRADECIMENTOS

Dedico este trabalho, realizado com todo amor, cuidado e responsabilidade, à meus pais que em nenhum momento da vida se opuseram a incentivar-me desde o Ensino Básico até hoje, mesmo atravessando momentos difíceis durante a vida.

À minha noiva, Marina Marques, pela paciência e entendimento do tempo consumido para realização do mestrado, das noites que deixamos de sair, dos dias que não nos víamos, além de momentos em que foi preciso suportar meus instantes de impaciência e nervosismo.

Aos meus avós, que não estão mais presentes, mas que de algum lugar observam o meu crescimento profissional.

Aos meus amigos que não só me apoiaram em situações difíceis, mas também proporcionaram momentos de alegria e descontração necessários para continuação do projeto.

À todos os professores que desde o ensino fundamental me trataram com o devido respeito e profissionalismo e puderam me ensinar, além da disciplina ministrada, ética e cidadania.

Ao meu amigo, professor e orientador da graduação Marcello Montillo Provenza por continuar me aconselhando, ensinando e apontando o caminho a seguir.

Aos professores José Francisco Pessanha, Vinicius Xavier, José Fabiano, Ricardo Accioly, Antônio Orestes, Julio Siqueira e tantos outros que mesmo após o término da graduação não refutaram em me auxiliar a qualquer momento.

À UERJ por me receber de braços abertos, como funcionário e aluno, propiciando um ambiente facilitador para o aprendizado e crescimento.

Aos meus orientadores, Fabiano Oliveira e Paulo Eustáquio, pelos ensinamentos, conselhos, cobranças e exigências, sempre de maneira tranquila, objetiva e bem orientada. Vocês foram essenciais para este projeto. Termino este trabalho com a certeza que conquistei uma das melhores orientações dentro do mestrado e que conquistei não só bons professores, mas sim amigos para o resto da vida.

Por fim, ao meu melhor amigo João Felipe, assassinado em 2017. Sempre de abraços abertos, sorriso no rosto, e de palavra confortante quando eu mais precisava. Sinto sua falta todos os dias. Foi difícil terminar sem sua presença. Por outro lado, veio de sua filha e minha afilhada uma das maiores forças para poder continuar e não desistir.

Obrigado à todos vocês, fizeram parte diretamente de todo o processo.

RESUMO

GOULART, Jorge Luiz de Jesus. *Avaliação de Usabilidade e Aprendizagem de Algoritmos com o uso do TuPy Online*. 2019. Dissertação (Mestrado em Ciências Computacionais) – Instituto de Matemática e Estatística, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2019.

Este estudo tem como objetivo avaliar a usabilidade e a eficácia da ferramenta TuPy Online na aprendizagem de disciplinas do curso de Ciência da Computação da Universidade do Estado do Rio de Janeiro (UERJ). O TuPy Online foi idealizado como uma ferramenta para visualização de execução de algoritmos, com ênfase na exibição de estruturas de dados em diferentes níveis de abstração. Este trabalho é o resultado da avaliação da aplicação sistemática da ferramenta em diversas turmas de Graduação e Pós-graduação. Utilizando a metodologia de Savi et al. para avaliação de aprendizagem, verificamos que a introdução da ferramenta produziu impacto positivo e, considerando o aspecto de percepção dos alunos, um impacto ainda mais expressivo. Ambas análises foram validadas com significância estatística. Foi também avaliada a usabilidade da ferramenta utilizando o questionário SUS. Além disso, aplicou-se um questionário em busca de um retorno específico referente às funcionalidades do TuPy Online. Ambos resultados foram positivos e propiciaram o mapeamento de possíveis melhorias do *software*.

Palavras-chave: Aprendizagem. Algoritmos. TuPy Online. Estatística Não-Paramétrica.

ABSTRACT

GOULART, Jorge Luiz de Jesus. *Usability Assessment and Algorithm Learning Using TuPy Online*. 2019. Dissertação (Mestrado em Ciências Computacionais) – Instituto de Matemática e Estatística, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2019.

This study aims to evaluate the usability and the effectiveness of the TuPy Online tool in the learning of Computer Science courses lectured at the State University of Rio de Janeiro (UERJ). TuPy was conceived as a tool for visualizing execution of algorithms, with an emphasis on displaying data structures at different levels of abstraction. This work is the result of the evaluation of the systematic application of the tool in several undergraduation and graduation classes. Using the methodology of Savi et al. for evaluation of learning, we verified that the introduction of the tool produced a positive impact and, if considered the perception aspect of the students, an even more expressive impact. Both analysis were validated with statistical significance. It was also done an usability evaluation of the tool. A usability evaluation of the tool was also made using the SUS questionnaire. In addition, a questionnaire was applied in search of a specific return regarding the functionality of TuPy Online. Both results were positive and provided the mapping of possible software improvements.

Keywords: Algorithms. Complexity of Algorithms. Empirical Analysis of Algorithms. Automatic Analysis of Algorithms.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

Nos últimos tempos, tem sido notória a migração de métodos tradicionais de ensino para métodos alternativos. Com o grande avanço tecnológico e a popularidade da Internet, ambientes digitais dão oportunidade a uma nova forma de aprendizagem, tornando-se frequente o desenvolvimento de *softwares* projetados para o complemento do processo de ensino.

> Por se tratar de uma civilização tecnificada e em permanente evolução, são exigidos cada vez mais conhecimentos e habilidades, em vários níveis, nos campos da cultura, ciência e tecnologia, criatividade e inovação. Tais qualidades cognitivas não são adquiridas apenas pela percepção passiva de informações, mas também pela adoção de processos educativos que as desenvolvam. [1].

No entanto, trabalhar com novas tecnologias no ensino não é trivial tanto quanto se poderia presumir, carecendo de preparação, habilidades básicas e materiais disponibilizados. Além disso, ao passo que alguns educadores empenham-se em ampliar o ensino e a aprendizagem com o uso de computadores, outros resistem à ideia do emprego de novas metodologias.

A criação de jogos e *softwares* educacionais, ambientes virtuais, ensino à distância, entre outros, não vieram acompanhados de metodologias adequadas para avaliar sua capacidade pedagógica. Entretanto, alguns pesquisadores têm buscado suprir essa deficiência concebendo e aprimorando alguns métodos de avaliação tais como: *EGameFlow* [2], o *MEEGA* [3], *TAM* [4], *PETESE* [5], *SALG* [6] que serão descritos melhor na próxima seção.

Em 2018, um grupo de pesquisadores da Universidade do Estado do Rio de Janeiro (UERJ), idealizou a ferramenta TuPy Online [7], para amparar e auxiliar estudantes na aprendizagem de algoritmos, e com três objetivos explícitos:

1. introduzir uma pseudolinguagem reduzindo o número de palavras-chave para construção de algoritmos;

2. basear-se na língua portuguesa com comandos e identificadores escritos em português, atendendo um público-alvo sem domínio da língua inglesa;

3. permitir ao usuário uma visualização descomplicada do progresso da estrutura de dados frente à depuração do código linha por linha.

A ferramenta TuPy Online foi desenvolvida como projeto de código-aberto, derivada da ferramenta Online Python Tutor, também de código-aberto. Tal ferramenta base implementou recursos para visualização de execução de programas originalmente escritos em Python, com visualização das variáveis como alocadas em memória. O TuPy Online permite customização da visualização em diversos níveis de abstração através do uso da linguagem DOT.

> O uso de descrições textuais na linguagem DOT, que possui uma extensa gama de opções para customização de formatos, das cores e do posicionamento dos elementos visuais, permite gerar visualizações parametrizadas pelo estado do programa. Como as funções de visualização geram programas DOT armazenados em cadeias de caracteres, é possível usar operações básicas como concatenação para manipular ou construir qualquer subprograma DOT válido, em tempo de execução. Isso significa que, além das funções predefinidas para estruturas de dados convencionais, o programador pode criar novas funções capazes de exibir imagens com as características que desejar, complementando a visualização de forma integrada. [7].

Após o desenvolvimento da ferramenta, ela foi experimentalmente utilizada em algumas turmas. Mesmo tendo havido grande aceitação da ferramenta por parte de alunos e professores, foi planejada uma avaliação estatística sobre sua eficácia na aprendizagem, bem como uma análise sistemática de usabilidade, para uma validação mais profunda. Neste trabalho, apresentamos o *software* TuPy Online e o estudo da avaliação de aprendizagem e de usabilidade com essa ferramenta, na Universidade de origem do mesmo.

Em relação a usabilidade, atualmente, há vários questionários padronizados para medição do ponto de vista do usuário e posterior à experimentação do *software*. Alguns exemplos de questionários bastante reconhecidos são: SUS [8], ASQ [9], PSQ [9], PSSUQ [9], CSUQ [9], entre outros.

Assim sendo, este trabalho tem dois objetivos básicos: o primeiro é discutir metodologias de avaliação de *software* do ponto de vista de ganhos de aprendizagem; decidir, adotar e definir, quais destas metodologias, inclusive testes estatísticos, conduzem a uma melhor avaliação do TuPy Online. Um segundo objetivo é estudar metodologias para avaliar a usabilidade, escolher e aplicar uma delas, de forma a propor melhorias no software, bem como no processo de sua introdução, visando a aprendizagem. Para tanto, foi utilizado um questionário padrão de avaliação de usabilidade e um outro questionário específico para o TuPy Online com o intuito de receber percepções detalhadas do uso desse *software* por parte dos alunos.

O trabalho está organizado da seguinte forma. No restante deste capítulo apresenta-se com mais detalhes o TuPy Online. No Capítulo 1, apresenta-se uma revisão da literatura de métodos de avaliação de aprendizagem e de usabilidade. No Capítulo 2, explicita-se a teoria sobre testes estatísticos que serão utilizados e a metodologia utilizada para análise da aprendizagem e de usabilidade obtida com o uso da ferramenta. No Capítulo 3, apresenta-se a aplicação da metodologia e uma discussão dos resultados. No Capítulo 3, conclui-se a análise assinalando pontos positivos e negativos da avaliação sobre a aprendizagem, usabilidade e do *feedback* dos alunos, além de propostas para trabalhos futuros.

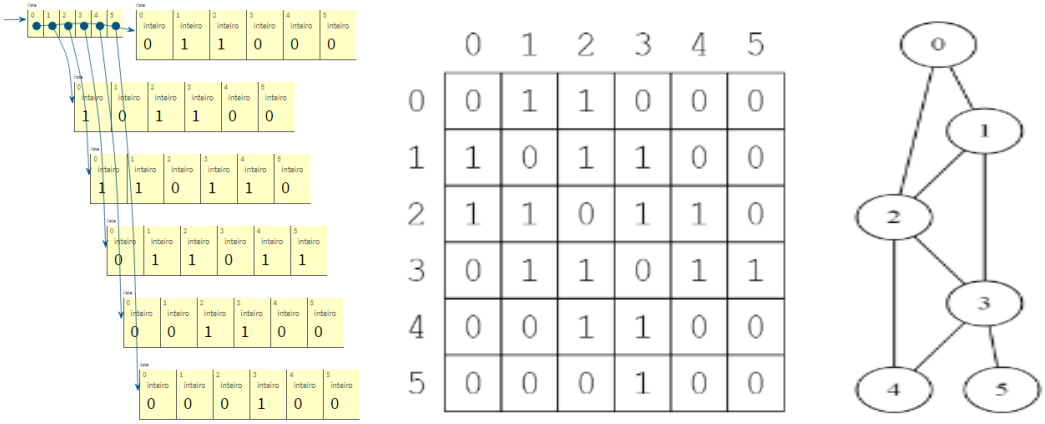
A ferramenta TuPy Online

A ferramenta TuPy Online foi criada para amparar e auxiliar estudantes na aprendizagem de algoritmos. Ela trabalha com pseudocódigos escritos em uma linguagem com estrutura semelhante a do Python, mas com todos os termos (palavras-chave e identificadores do algoritmo) em português. Permite acompanhar a execução do pseudocódigo passo a passo, mostrando o conteúdo das variáveis envolvidas e possibilita a visualização simultânea das estruturas de dados correspondentes.

A ferramenta foi criada no contexto de software livre e pode-se usá-la diretamente na rede ou a partir de uma cópia local. Em ambos os casos, a interface do TuPy Online é feita sobre o navegador de Internet.

Um objetivo mais ambicioso com a nova ferramenta foi o de permitir a construção de visualizações customizadas para sintetizar compreensivamente a abstração representada. A personalização é executada com o auxílio do *Graphviz* [10], originalmente criado para visualizações de grafos e diagramas estáticos, descritos na linguagem DOT [11]. A Figura 1 ilustra esse ponto.

Figura 1 – Visualização de uma matriz de adjacência de um grafo através de três formas no TuPy Online: a visualização tradicional do TuPy Online, a visualização com a função de abstração matriz, e com a função de abstração grafo.



Fonte: O próprio autor.

Arquitetura do TuPy Online

A ferramenta possui quatro componentes principais:

- *Online Python Tutor* OPT;
- Graphviz;
- ANTLR4;
- Interpretador para TuPy.

O OPT [1] é uma ferramenta de apoio ao aprendizado de Python e utilizada em diversas universidades em todo o mundo. O Graphviz [2] é um pacote de código aberto dedicado à representação de informação estruturada na forma de grafos e redes. Foi criado por pesquisadores da AT&T Labs em 1991 e possui apoio da comunidade até hoje. O ANTLR4 [3] é um gerador de analisadores sintáticos a partir de definições de gramáticas. É, por exemplo, usado no mecanismo de buscas do Twitter. Finalmente, o interpretador da linguagem TuPy foi desenvolvido em Python e é responsável pelo processamento do fluxo de execução do programa. Para o melhor entendimento da arquitetura, dois dos componentes são detalhados a seguir: o OPT e o interpretador da linguagem TuPy.

[1] Disponível em http://pythontutor.com/
[2] Disponível em https://graphviz.org. Acessado em 10 set. 2019.
[3] Disponível em https://www.antlr.org. Acessado em 10 set. 2019.

OPT (Online Python Tutor)

O OPT foi criado originalmente para a visualização de execução de programas escritos apenas em Python, mas, ao longo dos anos, recebeu suporte a novas linguagens. Um aspecto de sua arquitetura que favoreceu tal implementação é o fato da visualização não ser construída diretamente a partir da interpretação do programa Python, mas sim da interpretação de uma estrutura de representação intermediária, a qual é gerada a partir da execução íntegra do programa. A questão de estender o suporte a novas linguagens se dá, portanto, pela implementação de um mecanismo para geração dessa representação intermediária. Para as linguagens atualmente suportadas, existem componentes separados encarregados de executar o código com restrições de privilégios e implementar alguma estratégia para produzir a estrutura desejada.

A representação utilizada é um arquivo JSON que descreve o rastro da execução, isto é, contém uma sequência de informações acerca do estado de variáveis e da pilha de ativação para cada passo da execução do programa. Como essa execução acontece integralmente de forma não interativa, o usuário pode, em posse do arquivo retornado, navegar livremente pela visualização da execução, em passos para a frente ou para trás.

No OPT, o rastro da execução de um programa Python é obtido através da análise de dados fornecidos pelo bdb (módulo da linguagem Python para depurar códigos escritos na linguagem) ao executá-lo. No TuPy Online, a concepção de um interpretador permitiu a integração de um módulo que acompanha a execução dos programas e é capaz de produzir o arquivo de rastro de execução no formato desejado.

O Interpretador da Linguagem TuPy

Foi utilizada a ferramenta ANTLR4 para que, a partir de uma gramática EBNF[4], houvesse geração de código para cumprir as etapas de análise léxica e sintática do interpretador, gerando uma árvore de sintaxe abstrata (AST, abreviação do inglês *abstract syntax tree*) [12].

Cabe ao interpretador, em seguida, percorrer a AST e registrar a evolução de estado do programa. Para isso, foram implementadas abstrações para a tabela de símbolos e a pilha de ativação, que contêm as informações necessárias não somente à execução mas também à produção do rastro de execução.

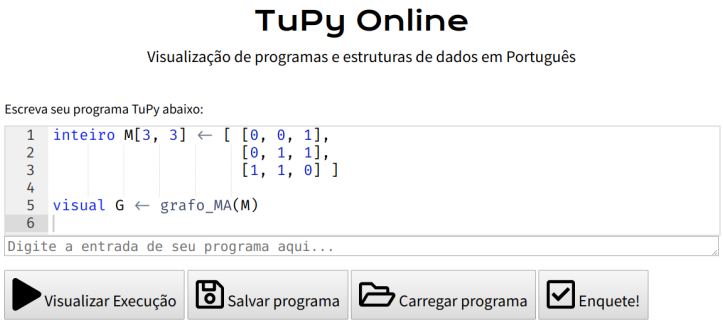
Considerando que tanto o serviço do lado do servidor do OPT quanto o interpretador TuPy foram desenvolvidos em Python, a comunicação entre os dois componentes pôde ser efetuada de forma direta.

Principais funcionalidades da ferramenta

A interface do TuPy Online é totalmente baseada no navegador e herda diversos aspectos do OPT. Conforme mostra a Figura 2, a página principal consiste do campo de edição de texto e dos botões de ação. A parte inferior da página também conta com um manual da linguagem TuPy e uma coleção de exemplos representativos de diferentes algoritmos e estruturas de dados. Durante a visualização, um painel de código situa-se à esquerda com destaque às linhas atual e seguinte da sequência de execução, junto aos botões de navegação. Assim como no OPT, é possível configurar pontos de parada por meio de um clique na linha de código desejada.

[4]Uma extensão do Formato Backus–Naur (BNF) que, dentre outras características, possui notações que facilitam opcionalidade e repetições em regras.

Figura 2 – A interface de edição de código do TuPy Online.

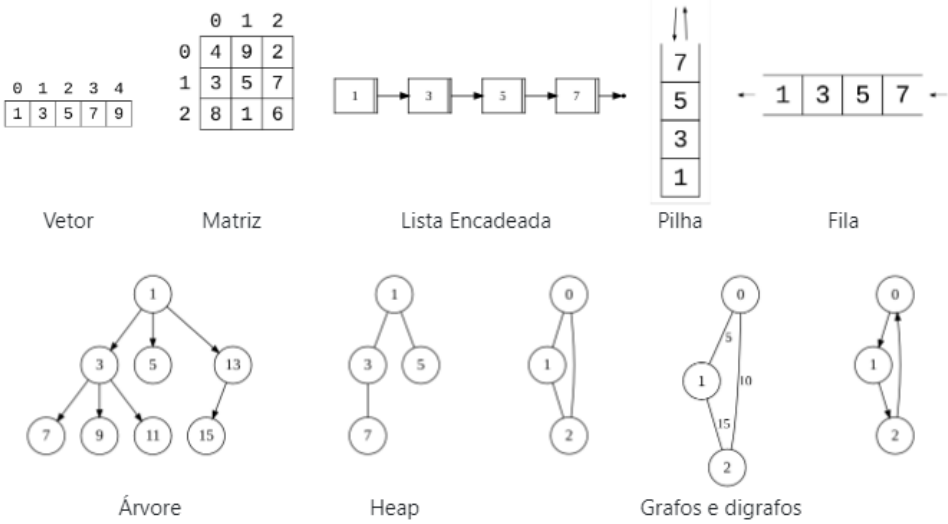


Fonte: O próprio autor.

Alguns pontos de destaque implementados são a possibilidade de fornecer dados de entrada do usuário a serem lidos pelo programa, assim como botões para salvar e carregar o código-fonte de um arquivo de texto. O editor de código também foi adaptado para habilitar conveniências como preenchimento automático, expansão ou contração de blocos (*code folding*) e uma fonte[5] com suporte a ligaduras voltadas à programação, isto é, a capacidade de unir símbolos consecutivos de um código em um novo símbolo coerente automaticamente. A digitação dos símbolos correspondentes ao comparador de não-igualdade de operandos "!=", por exemplo, resulta na exibição do símbolo "≠". Com isso, espera-se que o código TuPy possa se aproximar visualmente de notações de pseudocódigo manuscrito. Além disso, foi disponibilizada uma versão que pode ser executada localmente, sem necessidade de conexão com a internet, minimizando problemas de escalabilidade oriundos do compartilhamento de recursos do servidor.

Visualização de Estruturas de Dados

Figura 3 – Exemplos de visualizações para estruturas de dados no Tupy Online

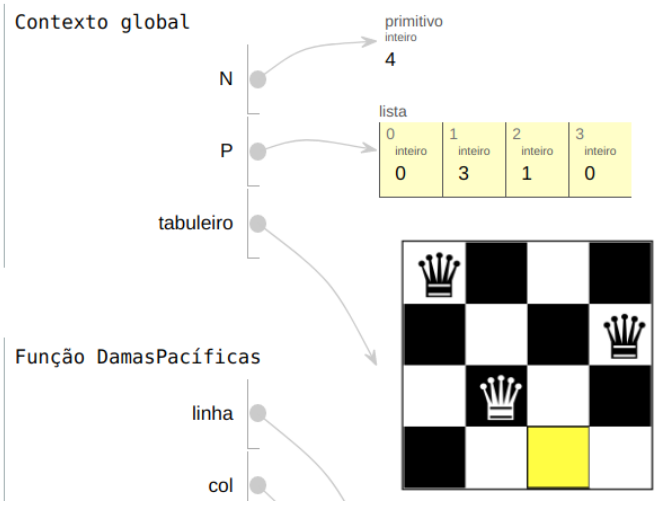


Fonte: O próprio autor.

[5]Fira Code. Disponível em https://github.com/tonsky/FiraCode. Acessado em 18 fev. 2018.

O projeto do TuPy Online providenciou uma biblioteca de funções para criar representações visuais extensíveis de estruturas de dados. A Figura 3 mostra as principais visualizações já prontas para diversas estruturas de dados: vetores, matrizes, listas encadeadas, pilhas, filas, árvores enraizadas, grafos e digrafos valorados ou não. Internamente, as funções encapsulam lógica para percorrer as estruturas convertendo-as para descrições textuais na linguagem DOT, que correspondem às respectivas representações gráficas. Tais descrições são consumidas e transformadas em imagens durante a execução no navegador do usuário, pelo Graphviz.

Figura 4 – Tabuleiro de xadrez gerado por uma função TuPy durante a execução do algoritmo de backtracking para o problema N Rainhas (N=4)



Fonte: O próprio autor.

Visualizações customizadas também podem ser criadas, como é o exemplo apresentado na Figura 4. Isto pode ser feito com o uso da linguagem DOT, que possui uma extensa gama de opções para a customização dos formatos, das cores e do posicionamento dos elementos visuais, o que permite gerar visualizações parametrizadas pelo estado do programa. Isso significa que, além das funções predefinidas para estruturas de dados convencionais, o programador pode criar novas funções capazes de exibir imagens com as características que desejar, complementando a visualização de forma integrada. Essa funcionalidade é observada na demonstração do problema das N rainhas em um tabuleiro de xadrez, incluído como um dos exemplos da ferramenta para ilustrar a customização da visualização para problemas específicos.

Exemplo da proximidade da visualização com a abstração da estrutura de dados

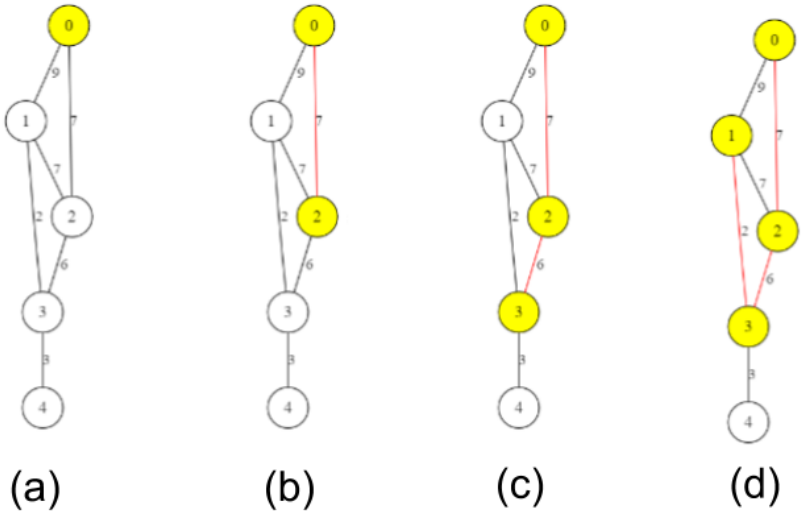
Considere o problema da obtenção de uma Árvore Geradora Mínima (AGM) em um grafo valorado. Essa é uma importante aplicação de grafos na interconexão de redes, pois uma AGM indica a forma mais barata de se obter interconexão dos diversos nós de uma rede, modelada como um grafo valorado.

Figura 5 – Representações em forma de listas (a), matrizes/vetores (b) e grafo (c)

Fonte: O próprio autor.

A Figura 5 ilustra três diferentes visualizações de uma AGM, obtida pelo algoritmo de Prim. Esse é um algoritmo guloso que constrói a AGM passo a passo, a partir de um vértice inicial qualquer. A cada passo um novo vértice e uma nova aresta do grafo original são escolhidos e integrados à árvore em construção. A aresta escolhida deve ser aquela de custo mínimo, sujeita à restrição de que um dos vértices incidentes já esteja na árvore e o outro não.

Figura 6 – Visão parcial da escolha das arestas: (b): aresta (0,2); (c): aresta (2,3) e (d): aresta (1,3).



Fonte: O próprio autor.

A parte esquerda da figura ilustra duas listas na memória. A da parte superior contém um conjunto de listas da representação interna da matriz de adjacências do grafo. A parte inferior mostra uma lista das arestas escolhidas: (2,0), (3,2), (1,3) e (3,4). Essas representações são aquelas feitas pelo OPT, bastante básicas. Na parte central da figura, o grafo é exibido como uma matriz de adjacências e as arestas escolhidas como um vetor. Pode-se notar que a visualização já torna mais fácil o entendimento do resultado.

Finalmente, na parte da direita, a representação é em forma gráfica usual de um grafo, muito mais próxima da abstração que normalmente se deseja sobre tal estrutura de dados. Aqui, as arestas escolhidas são coloridas de vermelho, permitindo comparar diretamente a sobreposição das duas estruturas: o grafo e a AGM. Observe que os vértices estão coloridos com amarelo. A possibilidade de colorir os vértices e as arestas permite uma melhor compreensão do desenrolar do algoritmo.

A Figura 6 ilustra que a visualização criada torna possível entender, passo a passo, o funcionamento do algoritmo, pois o grafo pode ser exibido colorindo-se de amarelo os vértices já escolhidos para a AGM até cada passo, deixando em branco aqueles ainda não escolhidos. Ao mesmo tempo, as arestas que fazem parte do AGM são destacadas, pois são pintadas de vermelho.

1 REVISÃO DE LITERATURA SOBRE AVALIAÇÃO DE APRENDIZAGEM E USABILIDADE

Neste capítulo apresenta-se uma revisão da literatura acerca das metodologias empregadas nas avaliações de ganho de aprendizagem em um contexto onde onde se utilizam ferramentas computacionais para alavancar o aprendizado. Além disso, apresentam-se, também, modelos frequentemente utilizados para avaliação da usabilidade de *software*. Todos os métodos descritos são baseados na aplicação de questionários aos alunos participantes.

Em linhas gerais, a grande maioria das criações de modelos analíticos de aprendizagem empenha-se em tratar um problema específico, escorando todo o planejamento do experimento para atender as características do ambiente de aprendizagem utilizado, seja ele um jogo, *software* educacional, ou outro método computacional. Por outro lado, em termos de usabilidade, há na literatura modelos bastante consolidados e empregados para tais avaliações.

Muitas das descrições de conceitos estatísticos apresentados nesta revisão de literatura, como por exemplo, alfa de Crobach, teste t pareado, entre outros, estão melhor descritos no Capítulo 2.

Inicialmente, apresentamos os conceitos teóricos de quatro autores que trataram de Aprendizagem em empresas e escolas. Estas são as principais referências encontradas nos estudos sobre introdução de ferramentas, que comentaremos adiante. São eles: Kirkpatrick [13], Keller [14], Bloom [15] e Sindre e Moody [16].

1.1 Os 4 níveis de Kirkpatrick

Proposto por Kirkpatrick [13], o modelo dos níveis de avaliação de treinamento é uma metodologia bastante influente entre empresas, setores de recursos humanos e áreas administrativas. Tal metodologia é utilizada para a avaliação e consequente classificação dos resultados de um treinamento em um determinado grupo de funcionários e/ou alunos. O modelo contempla a análise em 4 fatores: Reação, Aprendizado, Comportamento e Resultados (Tabela 1).

Tabela 1 – Os 4 níveis do modelo de avaliação de treinamento de Kirkpatrick.

Nível	Avaliação	Descrição e características	Exemplos de ferramentas e métodos
1	Reação	Avalia como os alunos se sentiram após o treinamento ou experiência de aprendizagem	*Happy-sheets*; formulários de feedback; reações verbais; questionários pós-treinamento.
2	Aprendizagem	Avalia o aumento de conhecimento ou capacidade	Avaliações e testes antes e depois do treinamento; entrevistas e observações.
3	Comportamento	Avalia os efeitos da nova aprendizagem no ambiente de trabalho	Observações e entrevistas ao longo do tempo para avaliar mudanças, relevância das mudanças, e sustentabilidade das mudanças.
4	Resultados	Avalia os efeitos do treinamento do aluno no negócio da empresa	Questionários pós-treinamento; observação como parte de um treinamento seqüencial e de *coaching* durante um período de tempo; medições de retrabalho, erros, etc., entrevistas com os participantes, seus gerentes e grupos de clientes.

Fonte: Savi et al..

Reação

O fator Reação é o primeiro nível na escala avaliativa. Identifica a percepção do aluno quanto ao conteúdo, ao objeto estudado, à experiência de aprendizagem e à importância para o seu desenvolvimento profissional. De acordo com Kirkpatrick, avaliar a reação é similar à análise da satisfação do aluno. É importante que os alunos reajam positivamente para uma avaliação efetiva da reação. Realizar a análise do nível 1 de Kirkpatrick é fácil, rápido e de baixo custo para o pesquisador.

Aprendizagem

O fator Aprendizagem é o segundo nível na escala. Definir este fator significa a mudança na forma de perceber a realidade e aumento de conhecimentos e habilidades. Ou seja, após avaliar a reação dos participantes tem-se que avaliar se de fato gerou-se aprendizado. Após traçados os objetivos de aprendizagem do treinamento estes indicadores devem ser utilizados na avaliação. Todos os participantes devem realizá-la através de um teste escrito, no caso de avaliação de conhecimento ou teste de desempenho, no caso de percepção da realidade. O ideal é que seja feita uma avaliação antes e outra posterior ao treinamento.

Comportamento

Kirkpatrick define o terceiro nível da seguinte forma: é a extensão da mudança de conduta que ocorre porque o aluno participou do treinamento. A mudança de comportamento só ocorre se a aprendizagem é efetiva. Ou seja, assim como a aprendizagem decorrerá da reação positiva do aluno, o comportamento obterá uma avaliação positiva quando a aprendizagem também obtiver. Após alcançarem os objetivos de aprendizagem do treinamento colocam-se em prática os novos conhecimentos e habilidades. Resumidamente, o objetivo de qualquer treinamento é que o aprendizado seja de fato posto em prática. Algumas formas avaliam a mudança de conduta tais como: listas de verificação de comportamento, análise de indicadores de desempenho de produtividade, questionários envolvendo seus professores ou colegas de classe.

Resultado

Este nível é o de maior importância em um treinamento dentro de uma empresa. É nesta última avaliação que verifica-se se, de fato, houve ganhos representativos para o negócio. Ou seja, obteve-se ganho porque os funcionários participaram do treinamento. Os resultados podem vir de: aumento de produção, melhor qualidade do produto, redução de custos, redução de acidentes, aumento de vendas, redução de rotatividade de pessoal, aumento do lucro ou retorno do investimento.

1.2 Modelo ARCS de Keller

O modelo *ARCS de Keller* [14] foi desenvolvido com a intenção de compreender de maneira mais eficaz os fatores que influenciam na motivação para aprender e resolver possíveis problemas motivacionais. O modelo resultante criado por Keller é uma síntese de 4 categorias de variáveis que englobam a maioria das áreas de pesquisa sobre motivação humana.

Normalmente, a motivação é altamente imprevisível e mutável sofrendo, às vezes, influência de fatores que fogem ao controle. Com isso, discute-se os limites da responsabilidade do professor em sala de aula. Em relação ao comportamento social, supõe-se que a motivação possa, então, ser controlada pela aplicação de regras e reforços. No entanto, quando se trata do ambiente escolar, a visão popular é que a motivação requer intuição e talento nativo [14]. Gerou-se então dois questionamentos: (i) É possível sintetizar as teorias de motivação humana em um modelo simples e de fácil utilização para o praticante?; (ii) É possível desenvolver uma abordagem sistemática oposta à intuitiva para projetar uma instrução motivacional?.

Criou-se, então, o *Modelo ARCS de Keller*. Sua metodologia é utilizada para melhorar o apelo motivacional de materiais instrucionais. O modelo também baseia-se na teoria do valor da expectativa e supõe que as pessoas são motivadas a se engajar em uma atividade, se há uma percepção de que necessidades pessoais possam ser atendidas (aspecto do valor) e se há uma expectativa positiva para o sucesso (expectativa). A categoria valor foi dividida por Keller em 2 fatores: Interesse e Relevância. A categoria expectativa foi mantida sem alteração. Uma nova categoria foi inserida: Resultados. Por fim, as 4 categorias receberam os nomes: Atenção, Relevância, Confiança e Satisfação. Estas categorias/condições precisam ser satisfeitas para que as pessoas se tornem e permaneçam motivadas. Tais categorias serão descritas de forma mais detalhada a seguir.

Atenção

É um pré-requisito para aprendizagem, pois a preocupação motivacional é a obtenção e manutenção da atenção. Ou seja, direcionar a atenção para estímulos apropriados. Há facilidade em ganhar atenção. O desafio de fato, é sustentar o ganho de atenção. O objetivo é encontrar o equilíbrio entre o tédio e indiferença versus hiperatividade e ansiedade. Dentre as 6 estratégias listadas na Tabela 2, as estratégias 5 e 6 são úteis para manutenção da atenção.

Tabela 2 – Obtenção e Manutenção da Atenção de acordo com Keller.

ATENÇÃO - ESTRATÉGIAS
1 - INCONGRUÊNCIA - CONFLITOS
Introduzir um fato que parece contradizer a experiência passada do aluno
Apresentar um exemplo que não parece exemplificar um dado conceito
Apresente 2 fatos igualmente plausíveis, dos quais um apenas pode ser verdadeiro
Defenda o fato falso
2 - MATERIALIDADE
Mostre representações visuais de um conjunto de ideias ou relacionamentos
De exemplos de todo conceito instrucionalmente importante
Use anedotas relacionadas ao conteúdo, estudos de caso, biografias, etc.
3 - VARIABILIDADE
Varie o tom da voz e use movimentos do corpo, pausa e adereços
Varie o formato da instrução de acordo com o período de atenção do público
Varie o meio de instrução (entrega de plataforma, filme, vídeo, impressão, etc.)
Divida os materiais impressos usando espaços em branco, elementos visuais, tabelas, diferentes tipos de letra, etc.
Altere o estilo de apresentação (humorístico-sério, rápido-lento, alto-suave, ativo-passivo, etc.)
Alterne entre a interação aluno-instrutor e a interação aluno-aluno
4 - HUMOR
Quando apropriado, use peças de palavras durante a apresentação de informações redundantes
Use apresentações humorísticas
Use analogias humorísticas para explicar e resumir
5 - QUESTIONAMENTOS
Use técnicas de criatividade para que os alunos criem analogias e associações incomuns ao conteúdo
Crie atividades de resolução de problemas em intervalos regulares
Dê aos alunos a oportunidade de selecionar tópicos, projetos e tarefas que atraem sua curiosidade e precisam ser explorados
6 - PARTICIPAÇÃO
Use jogos, dramatizações ou simulações que exigem participação do aluno

Fonte: O próprio autor.

Relevância

Nesta categoria, a importância do conteúdo para o aluno é o foco. Para isso, alguns professores e instrutores de cursos tentam fazer com que o aprendizado seja relevante para oportunidades de carreira para os alunos (estratégias 2 e 3 da Tabela 3). Outros acreditam que o aprendizado deve ser, por si só, algo que seja desfrutado e valorizado. E, em um terceiro aspecto, a relevância pode surgir da forma como aquele conhecimento é ensinado, e não do conteúdo em si (estratégias 4 e 5 da Tabela 3).

Tabela 3 – Relevância de acordo com Keller.

RELEVÂNCIA - ESTRATÉGIAS
1 - EXPERIÊNCIA
Declare explicitamente como a instrução se baseia nas habilidades existentes do aluno
Use analogias familiares ao aluno de experiências anteriores
Descubra quais são os interesses dos alunos e relacione-os à instrução
2 - VALOR PRESENTE
Declare explicitamente o valor intrínseco em aprender o conteúdo, diferentemente de seu valor como um caminho para objetivos futuros
3 - UTILIDADE FUTURA
Declare explicitamente como a instrução se relaciona com as atividades futuras do aluno
Peça aos alunos para relacionar a instrução aos seus próprios objetivos futuros
4 - CORRESPONDÊNCIA
Para melhorar o comportamento de conquista de esforços, ofereça oportunidades para alcançar excelência sob condições de risco moderado
Para tornar a instrução sensível ao motivo do poder, ofereça oportunidades de responsabilidade, autoridade e influência interpessoal
Para satisfazer a necessidade de afiliação, estabeleça confiança e ofereça oportunidades para interação cooperativa sem risco
5 - MODELAGEM
Traga alunos do curso como palestrantes convidados entusiastas
Em um curso individualizado, use aqueles que terminarem primeiro como tutores
Entusiasmo pelo assunto ensinado
6 - ESCOLHA
Fornecer métodos alternativos significativos para atingir um objetivo
Forneça escolhas pessoais para organizar o trabalho

Fonte: O próprio autor.

Confiança

O terceiro componente, Confiança, pode influenciar a persistência e realização do aluno. Existem muitos fatores que contribuem para o nível de confiança ou expectativa de sucesso. Pessoas confiantes atribuem o sucesso à habilidade e ao esforço, ou seja, tendem a acreditar que podem alcançar seus objetivos por meio das próprias ações. Por outro lado, pessoas menos confiantes preocupam-se mais com o fracasso do que o próprio professor percebe. Neste caso, o envolvimento com o ego e a necessidade de impressionar os outros tornam-se foco. Assim, Keller criou a Tabela 4 para Confiança, com estratégias para ajudar o aluno a entender que a obtenção de algum nível de sucesso é resultante de algum esforço próprio.

Tabela 4 – Confiança de acordo com Keller.

CONFIANÇA - ESTRATÉGIAS
1 - REQUISITOS DE APRENDIZAGEM
Incorporar objetivos de aprendizado claramente declarados e atraentes em materiais instrucionais
Fornecer ferramentas de autoavaliação baseadas em objetivos claramente definidos
Explique os critérios para avaliação de desempenho
2 - DIFICULDADE
Organize materiais em um nível crescente de dificuldade; isto é, estruture o material de aprendizagem para fornecer um desafio "conquistável"
3 - EXPECTATIVAS
Inclua declarações sobre a probabilidade de sucesso com determinados esforços e habilidades
Ensine os alunos a desenvolver um plano de trabalho que resultará na realização de metas
Ajude os alunos a definir metas realistas
4 - ATRIBUIÇÕES
Atribua o sucesso do aluno ao esforço, em vez de sorte ou facilidade de tarefa, quando apropriado
Encoraje os esforços dos alunos para verbalizar as atribuições apropriadas tanto para os sucessos como para os fracassos
5 - AUTO-CONFIANÇA
Permita que os alunos se tornem cada vez mais independentes para aprender e praticar uma habilidade
Peça aos alunos que aprendam novas habilidades em condições de baixo risco, mas pratiquem o desempenho de tarefas aprendidas em condições realistas
Ajude os alunos a entenderem que a busca da excelência não significa que qualquer coisa aquém da perfeição seja o fracasso

Fonte: O próprio autor.

Satisfação

Esta categoria incorpora pesquisas e práticas que ajudam as pessoas a se sentirem bem com suas realizações. As pessoas devem ser mais motivadas se a tarefa e a recompensa forem definidas. No entanto, há de se ter o cuidado com pessoas que possam ficar ressentidas quando lhe dizem o que é preciso ser feito e a respectiva recompensa. Há maneiras

adequadas para o uso de recompensas em situações de aprendizagem que estimulem a recompensa instrucional. A Tabela 5 demonstra tais maneiras.

Tabela 5 – Satisfação de acordo com Keller.

SATISFAÇÃO - ESTRATÉGIAS
1 - CONSEQUÊNCIAS NATURAIS
Permitir que um aluno use uma habilidade recém-adquirida em um cenário realista o mais rápido possível
Aumentar verbalmente o orgulho intrínseco de um aluno em realizar uma tarefa difícil
Permitir que um aluno que domine uma tarefa ajude outras pessoas que ainda não o fizeram
2 - RECOMPENSAS INESPERADAS
Recompense o desempenho da tarefa intrinsecamente interessante com recompensas inesperadas e não contingentes
Recompense tarefas chatas com recompensas extrínsecas e antecipadas
3 - RESULTADOS POSITIVOS
Dê valor ao progresso ou realização bem-sucedida
Dê atenção pessoal aos alunos
Fornecer feedback informativo e útil quando for imediatamente útil
Elogiar imediatamente após o desempenho da tarefa.
4 - INFLUÊNCIAS NEGATIVAS
Evite o uso de ameaças como meio de obter desempenho de tarefas
Evite a vigilância (em oposição à atenção positiva)
Evite avaliações de desempenho externas e sempre que for possível ajudar o aluno a avaliar seu próprio trabalho
5 - AGENDAMENTOS
Forneça reforços frequentes quando um aluno estiver aprendendo uma nova tarefa
Fornecer reforço intermitente à medida que o aluno se torna mais competente em uma tarefa
Varie o cronograma de reforços em termos de intervalo e quantidade

Fonte: O próprio autor.

Resumidamente, essas 4 categorias formam a base do modelo *ARCS* com objetivo de compreender os fatores que influenciam no teor motivacional.

1.3 Taxonomia de Bloom

A Taxonomia de Bloom [15], também conhecida como Taxonomia de Objetivos Educacionais, é uma estrutura para classificar as declarações do que esperamos ou pretendemos aprender como resultado da instrução. Foi concebida como meio facilitador da troca de itens de teste entre professores de instituições diferentes, a partir da criação de um banco de dados, medindo o mesmo objetivo educacional. De acordo com Bloom, a ferramenta era mais do que um método de medição. A metodologia poderia ser uma linguagem comum sobre objetivos de aprendizagem, um facilitador entre pessoas, temas e níveis de ensino, ou como base para determinar metas educacionais amplas de um determinado curso ou currículo.

A Taxonomia forneceu definições para cada uma das 6 categorias do domínio cognitivo: Conhecimento, Compreensão, Aplicação, Análise, Síntese e Avaliação. Todas as categorias foram divididas em subcategorias, exceto a categoria Aplicação. Uma evidência da grande importância da Taxonomia de Bloom, segundo Krathwohl [17] é a sua citação e tradução para 22 idiomas.

A Taxonomia de Objetivos Educacionais fornece uma estrutura organizacional compreendendo objetivos classificados em cada uma de suas categorias. Enfatiza-se também a avaliação de aprendizagem com muitos exemplos de itens de teste fornecidos para cada categoria. Usar a Taxonomia para classificar objetivos, atividades e avaliações fornece uma representação visual clara e concisa de um determinado treinamento.

1.4 Aprendizagem de curto termo e longo termo

De acordo com Sindre e Moody [16], apesar de exames finais e pesquisas de avaliação de fim de semestre poderem auxiliar na avaliação de ganho de aprendizagem, não são

projetados para este propósito, e há problemas inerentes em usá-los para tal fim.

Com isso, criou-se uma avaliação de aprendizagem baseada em metas de aprendizagem da disciplina em curto termo e no contexto do programa educativo global e da vida profissional futura (longo termo).

Os autores também afirmam a importância em estabelecer se a mudança em um determinado curso foi bem sucedida na melhoria do aprendizado do aluno. Os autores sugerem evitar julgamentos tendenciosos dos professores suscetíveis a vieses cognitivos - o professor pode buscar evidências que confirmem sua tese, evitando evidências negativas. Além disso, variáveis de confusão podem estar presentes e enviesar o estudo tais como: (i) diferença na característica dos alunos; (ii) utilização da mesma avaliação em períodos consecutivos; (iii) avaliadores tendenciosos; (iv) tendência na normalização das notas; (v) mudanças no resultado ao se aplicar seguidamente o método.

A teoria aplicada [16] dá importância aos objetivos de aprendizagem que guiam a seleção de métodos de ensino e atividades para alcance de tais objetivos. Os autores citam 3 diferentes objetivos de aprendizagem:

- Conhecimento: Conceitos que o aluno deve entender.
- Habilidades: Quais tarefas os alunos são capazes de realizar.
- Atitudes: Atitudes, crenças e motivação que os participantes devem possuir.

A aprendizagem é um processo contínuo. Enquanto ela pode ser avaliada dentro do contexto dos objetivos de uma determinada disciplina, ela também pode ser avaliada em um contexto mais amplo. Com isso, os autores apresentam duas visões de aprendizagem:

- Aprendizagem de curto termo: se o curso foi bem sucedido em alcançar os objetivos de aprendizagem declarados.
- Aprendizagem de longo termo: se o curso contribuiu para o aprendizado geral do aluno.

Para operacionalizar o modelo acima, os itens da pesquisa precisam ser desenvolvidos para medir cada uma das aprendizagens, tanto de longo termo, quanto de curto termo.

Já em 2003, os autores [16] demonstravam-se surpresos com a não padronização na literatura, de um modelo que pudesse avaliar o ganho de aprendizagem com novos tipos de metodologia de ensino. Até então, existiam poucos instrumentos de qualquer tipo para avaliações de ganho de conhecimento. Os autores citam que as instituições desenvolviam instrumentos próprios de medição exclusivamente para atender seus próprios propósitos. Resumidamente, só encontraram dois modelos de avaliação: *The Student Assessment of Learning Gains (SALG)* [6] e *Student Opinion Survey of the Learning Process* [18].

O primeiro, *SALG*, foi desenvolvido para abordar as limitações dos levantamentos tradicionais de avaliação de um curso. A instrumentação da metodologia é baseada na análise de respostas qualitativas dos participantes, consistindo em um conjunto de perguntas que podem ser adaptadas para um determinado curso. No entanto, não há testes de confiabilidade e validação para o *SALG*, apenas o *feedback* positivo.

O segundo, *Student Opinion Survey of the Learning Process*, foi desenvolvido com o mesmo propósito do *SALG*. No entanto, tem como diferença o fato de ser mais geral. Enquanto o *SALG* era completamente adaptável, o modelo desenvolvido por Snare não precisava desta adaptação, ou seja, era por si só, aplicável a qualquer curso. No entanto,

nenhuma validação empírica é relatada para esta metodologia. Além disso, os dois modelos definem um conjunto de itens de pesquisa *ad-hoc*. Resumidamente, podem não medir, de fato, quaisquer construções teóricas subjacentes.

1.5 Estudos de Caso sobre Avaliação de Aprendizagem

Em 2009 [19], propôs-se um modelo de avaliação de ganho de aprendizagem e usabilidade de *software*, com metodologia arquitetada sobre padrões internacionais da ISO/IEC 9126. Basicamente, este modelo é configurado em duas partes: a primeira, voltada para fatores de qualidade externa e interna do *software* (Funcionalidade, Confiabilidade, Usabilidade, Eficiência, Sustentabilidade e Portabilidade) e a segunda, orientada para fatores de qualidade em uso (Eficácia, Produtividade, Segurança e Satisfação). Os autores revelam que as abordagens para avaliação levam em conta tanto o processo quanto o resultado final do produto. Além disso, assim como Sindre e Moody [16], corroboram a importância de uniformizar critérios, abordagens e métodos específicos para análise que, uma vez selecionados, devem ser adaptados ao ambiente em questão. O modelo proposto possui 27 sub-características a serem avaliadas pelos aspectos técnicos, educacionais, organizacionais e avaliação de conteúdo. Por fim, essas sub-características são agrupadas em 6 características, a saber: Funcionalidade, Confiabilidade, Usabilidade, Eficiência, Manutenção e Portabilidade.

Também em 2009, pesquisadores de Taiwan idealizaram o *EGameFlow* [2], outra metodologia de avaliação de *software* educacional. Os autores perceberam e relataram a importância do teor motivacional correlato à eficácia da aprendizagem do aluno, ou seja, o prazer motiva o estudo e o trabalho continuado, tornando-se peça chave para a avaliação. Resumidamente, a instrumentação do *EGameFlow* restringe-se a uma avaliação, por lista de verificação para cada um dos oito fatores considerados (Concentração, Clareza do Objetivo, Comentários, Desafio, Autonomia, Imersão, Interação Social, Melhoria do Conhecimento), principalmente avaliando o grau de prazer do usuário. A metodologia do *EGameFlow* é baseada na teoria do desenvolvimento de escalas de avaliação [20]. O estudo foi realizado com alunos do curso on-line "Introdução às Aplicações de *Software*" em uma universidade nacional do norte de Taiwan. Após o levantamento de dados, foram utilizados testes estatísticos bem conhecidos: alfa de Cronbach [21], ANOVA e teste-t [22]. Por fim, o questionário foi validado com 42 questões dispersas em 8 fatores.

No estudo de Savi [3], propõe-se o *MEEGA*, uma metodologia que visa ser de rápida e fácil aplicação, não requerendo ao aplicador competências avançadas na área de educação, medição e estatística. Ela é baseada em métodos de avaliação fundamentados em instrumentações padronizadas, certificadas por análises de validação e confiabilidade. A abordagem é apoiada no modelo de 4 níveis de Kirkpatrick. Desses 4, utilizou-se apenas o nível 1, Reação, dividido em três pilares (Motivação, Experiência e Aprendizagem), em que a reação do aluno determinará a qualidade do *software*. O pilar motivacional foi subdividido em 4 componentes (Atenção, Relevância, Confiança e Satisfação), de acordo com o modelo ARCS *de Keller*. Já o pilar Aprendizagem é estruturado em Conhecimento, Compreensão e Aplicação que são os 3 níveis iniciais da Taxonomia de Bloom além da inclusão de 2 outros fatores: Aprendizagem de Longo Termo e Aprendizagem de Curto Termo. É usado um primeiro questionário, que possui 27 questões e é baseado na escala Likert de 5 pontos, variando de -2 (discordo fortemente) até 2 (concordo fortemente). Já um segundo questionário, analisa o Conhecimento, a Compreensão e a Aplicação dos conceitos da disciplina através de uma auto-avaliação pelos alunos. Esta metodologia foi utilizada para analisar 3 jogos educacionais, e em seguida, avaliou-se o coeficiente alfa

de Cronbach. De acordo com Savi [23], o questionário foi validado com confiabilidade estatística.

Uma outra abordagem encontrada [5], denominada *PETESE*, para avaliação de *softwares* educacionais com referenciais técnicos (usabilidade) e pedagógicos (aprendizagem). Os autores citam a raridade de ferramentas para análise com objetivos educacionais que respeitem a preservação de padrões de qualidade. Esta abordagem utiliza-se de uma metologia análise/síntese que caracteriza o processo geral da elaboração do modelo em 4 etapas (Identificação, Análise, Síntese e Validação), inspirado nas obras de Silvern [19]. O questionário possui 69 questões divididas em alguns sub-itens. Para validação da metodologia foi utilizado o *software* educativo de matemática GGBook criado pelo laboratório ABACO da Universidade de Brasília. O questionário possui pouca flexibilidade para adaptação em outras situações. Apesar disso, o estudo foi conclusivo para o *software* em questão.

Em um estudo na Universidade de Santa Catarina [24], implementou-se uma revisão sistemática de literatura para a discussão de métodos de avaliações de jogos educacionais, considerando, principalmente, fatores como: aspecto de análise avaliado, projeto da pesquisa, modelos de observação, sistemática empregada, variedades de instrumentação de coleta de dados, dimensionamento de amostras e replicações, etc. Além disso, menciona-se a importância de cada procedimento adotado no experimento (escopo, planejamento, operação, análise e interpretação e apresentação de resultados), bem como o emprego de métodos estatísticos previamente validados por estudos empíricos. Nesta revisão, percebeu-se que o fator de avaliação mais frequente é o ganho de aprendizagem através de melhoria de competência aplicando um pré-teste e pós-teste ou uma autoavaliação, através de estudos de caso. Da mesma maneira, a experiência do usuário e a usabilidade baseada em termos de facilidade de uso, capacidade de aprendizado, utilidade, visualização agradável, eficiência, intuitividade e interação foram outros fatores de análise fortemente considerados em tais pesquisas. No entanto, 81% das avaliações não utilizaram nenhum método bem definido para realizar a avaliação. Além disso, a condução da pesquisa feita por metodologias *ad-hoc* é frequentemente encontrada. Alguns poucos estudos utilizaram questionários padrões para avaliações bem definidas, como o *MEEGA* [3], EGameFlow [2], TAM [4] ou SUS [8], que têm sido sistematicamente desenvolvidos e validados estatisticamente. Para análise de dados, tanto testes paramétricos (média como medida central - teste t, ANOVA, teste f, teste z), bem como testes não paramétricos (mediana como medida central - Mann-Whitney, Wilcoxon, Kruskal-Wallis) são empregados.

Recentemente, a metodologia do *MEEGA* foi adaptada para a elaboração do *MAAVA* [25] - Modelo de Avaliação de Ambientes Virtuais - com a inclusão do nível 2 do modelo de avaliação de Kirkpatrick. Este nível analisa o ganho de aprendizagem no treinamento através de dois testes: um pré-teste e um pós-teste. Além disso, a autora propõe a exclusão de alguns itens do primeiro questionário de Savi et al. a fim de tornar o questionário menos exaustivo. Foram também retiradas questões desnecessárias para avaliação de ambientes virtuais. O *software* avaliado foi o *Educ-MAS GA* na disciplina de Geometria Analítica do curso de Engenharia de uma universidade pública do Rio de Janeiro. O experimento foi realizado em apenas 3 dias. No primeiro dia, foram aplicados o pré-teste e o questionário de caracterização; no segundo, a realização do treinamento com a utilização do *software*; no terceiro, o pós-teste. O estudo foi conclusivo, e relatou o ganho de aprendizagem dos alunos. Segundo a autora, o número pequeno de alunos e um curto período de experimentação são limitações da análise.

Uma pedagogia de aprendizagem ativa é uma aprendizagem baseada em equipe, que

possui bom desempenho em 4 vertentes: desempenho do aluno, trabalho em equipe, pensamento crítico e satisfação com a aprendizagem. Para a imersão em ambientes educacionais de farmácia [26] alguns *softwares* e processos já foram utilizados, como por exemplo, o *Second Life* e outros *softwares* de realidade virtual (VR). O *Second Life* é um ambiente virtual e tridimensional que simula a vida real e que pode ser encarado como um jogo, um mero simulador, um comércio virtual ou uma rede social. Neste modelo de avaliação [26], 18 estudantes participaram da análise, divididos em grupos randomizados pelos próprios alunos. A participação foi voluntária sem quaisquer benefícios extras. A realização do experimento, em um único dia, foi da seguinte forma: um exercício de leitura, um teste de garantia de prontidão inicial, um teste de garantia de prontidão de equipe e um exercício de aplicação *VR*. O questionário de avaliação foi composto por 14 questões na escala Likert além de 2 questões abertas. No entanto, o estudo possui algumas limitações como o possível interesse prévio dos participantes por *VR*, o tamanho da amostra, e pelo fato do questionário considerar unicamente a perspectiva do usuário sem avaliar o seu desempenho real posterior à experimentação. O estudo foi conclusivo e os autores relataram que os estudantes sentiram que aprenderam mais no ambiente de *VR* e que o sistema oferecia ferramentas adequadas para o ensino.

Outro mecanismo para análise de jogos educacionais é uma abordagem baseada na externalização da avaliação [27], ou seja, oferece a possibilidade de alterar a lógica avaliativa sem sequer modificar os mecanismos do jogo. Neste trabalho foi criado o *EngAGe* (mecanismo de avaliação em jogos), recuperando informações sobre os *gameplays* através de um painel de *LA (Learning Analytics)*. O modelo é composto por 6 fases: Seleção, onde os professores relacionam quais dados são relevantes para o estudo (tempo, contagens, interação do jogo, pontuações e dados do jogador); Coleção, onde todos os dados dos alunos de cada professor são coletados; Análise, ocasião em que os dados recolhidos são agregados e analisados por técnicas de mineração de dados; Visualização, que exibe tempo consumido em jogo e pontuação média acompanhado de anomalias detectadas e previsões, fazendo uso de variáveis instrumentais; Interação, onde o professor interage com a informação; Reação, na qual o professor age adaptando o jogo para o propósito particular em questão. Além disso, o processo utiliza-se de métodos de análise de agrupamentos para avaliar os jogadores, como por exemplo, o método *k-means*. A avaliação do *EngAGe* com educadores foi composta por um pré e um pós-teste (uso da escala Likert de 5 pontos, variando entre "muito difícil de usar" e "muito fácil de usar", e, em alguns casos, entre "menos útil" e "muito útil") de um modelo pré-experimental sem grupo de controle. Os dados coletados no estudo foram analisados utilizando estatística descritiva e não-paramétrica. Sob a perspectiva de usabilidade, o questionário empregado foi o *SUS* [8], que provou ter versatilidade e robustez. O estudo resultou em uma avaliação positiva com valores globais positivos. Tanto o editor de avaliação quando o *LA* foram classificados como úteis pelo participante.

Para avaliações de *CVE's* (ambientes virtuais colaborativos multimodais) propôs-se uma versão adaptada do modelo pedagógico P2 [28]. O autor [29] sugere a criação deste modelo e o utiliza em um estudo experimental com professores e alunos do ensino médio, entre 11 e 18 anos, para medir o impacto da imersão no ganho de aprendizagem do estudante. O ambiente virtual avaliado remonta o Parlamento Europeu em Bruxelas e imerge os alunos em debates online. Após a avaliação, compara-se o desempenho entre o ambiente imersivo e o não-imersivo (ex: sala de aula). O modelo discute ainda a percepção das 3 dimensões de imersividade: Imersão espacial, espaço em que alunos e professores possam relacionar-se criando objetivos individuais; Imersão emocional, relação sentimental entre os participantes da atividade; e imersão temporal, a agregação de aprendizagem conduz e

estimula o interesse dos alunos. A medição para avaliação seguiu os seguinte critérios: (i) eficácia, tarefas concluídas e taxas de erro; (ii) eficiência, tempo necessário de execução de tarefas; (iii) satisfação, utilizando o Questionário de Usabilidade do Sistema Pós-Estudo (PSSUQ) [30]; (iv) satisfação do usuário através de um questionário padronizado [31]; (v) e o ganho de aprendizagem através do comportamento argumentativo [32]. Este estudo também fez uso do alfa de Cronbach para concluir validação e a consistência interna dos questionários utilizados. O estudo representou um retorno positivo para o *software* estudado, principalmente resultando em um impacto positivo na geração de ideias entre os alunos. Também permitiu que os participantes criassem um espaço (imersão espacial) para que a narrativa de aprendizado da tarefa do grupo ocorresse.

Outra metodologia [33] criada para verificação de ganhos e retenção de aprendizagem para modelos de simulação 2D e 3D ressalta a importância da fidelidade da simulação ao mundo real. Enquanto a baixa fidelidade facilita a abstração por parte do aluno, consequentemente estimula o seu aprendizado, a alta, apesar de apresentar baixa motivacional para estudantes novatos, auxilia a transferência (Aplicação da aprendizagem no mundo real) de conhecimento para resolução de problemas reais (principalmente na engenharia). O estudo também revela que a capacidade cognitiva do participante também influencia no processo de aprendizagem. Pessoas com menor capacidade podem apresentar ganhos maiores utilizando novas metodologias. É importante destacar também, a escassez da literatura fora dos meios de educação referente a retenção de conhecimento em ambientes simulados assertivamente discutido pelo autor. O questionário da ferramenta utilizou uma escala Likert de 5 pontos, avaliações pré e pós-teste, além de avaliar o tempo de execução de tarefa e o grau de erros cometidos (caso houvesse). A avaliação durou entre 2 e 4 semanas, com 63 participantes, e para análise foram utilizados testes em ANOVA.

Em nenhum dos experimentos comentados foi inserido grupo de controle. Grupo de controle é parte vital de métodos científicos, permitindo o estudo experimental de uma variável por vez. De acordo com Montgomery [34], a intenção da introdução do grupo de controle é comparar um determinado grupo sob o efeito de um tratamento com outro grupo (grupo de controle) sem tratamento.

1.6 Metodologias para Avaliação de Usabilidade de *Software*

Nesta Seção, abordaremos com detalhes as avaliações de usabilidade. Usabilidade é uma qualidade caracterizada pela facilidade em usar um sistema iterativo, ou seja, é a eficaz relação entre tarefa, usuário, interface e demais aspectos do *software* utilizado [35]. Uma das formas de medir a usabilidade de um *software* é através de um teste de usabilidade. Tal teste é uma maneira sistemática para observar realmente a utilização experimental do programa pelo usuário [36]. O objetivo principal deste tipo de teste é verificar e corrigir possíveis problemas de usabilidade no produto antes da liberação [37]. Em geral, a maneira mais comum de utilizar tal teste de usabilidade a partir do ponto de vista do usuário é através de questionários padronizados.

De acordo com Nielsen [38], a usabilidade é um atributo qualitativo para determinar a facilidade em usar as interfaces do programa. A usabilidade, para Nielsen [38], é subdividida em 5 componentes de qualidade:

- **Capacidade de aprendizado**: facilidade em realizar uma determinada tarefa no primeiro contato com o *software*;
- **Eficiência**: Execução rápida de tarefas após o conhecimento prévio do *software*;

- **Memorização**: Após um período de não utilização do *software*, facilidade em restabelecer a capacidade do usuário;
- **Erros**: Quantos erros são cometidos pelo usuário, qual o grau de gravidade e se são recuperados facilmente;
- **Satisfação**: A interface é agradável.

Em 1995, Lewis [9] descreveu pesquisas realizadas sobre usabilidade pela IBM. O foco desta pesquisa foi a aplicação de métodos psicométricos ao desenvolvimento de questionários para medir a satisfação do usuário com a usabilidade do sistema. Além disso, a pesquisa fornece 4 questionários da IBM, discutindo as instruções de administração e pontuação dos testes. Os 4 questionários são os: *The After Scenario Questionnaire* (ASQ), *The Printer Scenario Questionnaire* (PSQ), *The Post-Study System Usability Questionnaire* (PSSUQ), *The Computer System Usability Questionnaire* (CSUQ).

> Most usability evaluations gather both subjective and objective quantitative data in the context of realistic scenarios-of-use, as well as descriptions of the problems representative participants have trying to complete the scenarios. Subjective data are measures of participants' opinions or attitudes concerning their perception of usability. Objective data are measures of participants' performance (such as scenario completion time and successful scenario completion rate). [9].

O ASQ é um questionário de 3 itens que os avaliadores de usabilidade da IBM usaram para avaliar a satisfação do usuário após a conclusão do experimento. Os itens abordam 3 importantes componentes: facilidade de conclusão de tarefa, tempo para concluir uma tarefa e adequação das informações de suporte. É um questionário simples, leva pouco tempo para os participantes concluírem - consideração prática importante para os estudos de usabilidade. As 3 questões são avaliadas em escalas gráficas de 7 pontos, ancoradas nos pontos finais com os tempos "Concordo plenamente" em 1, e "Discordo plenamente" em 7. Por fim, ao ser empregado como um avaliador de usabilidade, o ASQ obteve evidências que o indicam como confiável, sensível e válido.

O PSQ é uma versão anterior do ASQ. Trata-se de um questionário de satisfação do participante construído para uma série de estudos sobre impressoras entre 1983 e 1984. No PSQ os itens são escalas de 5 pontos, ancoradas nos pontos finais com os termos "Aceitável como está" em 1, e "Precisa de muita melhoria" em 5. Os resultados de um estudo teste indicaram que o PSQ é uma medida razoavelmente sensível, ou seja, é uma medida que pode sofrer influência de valores extremos no questionário com certa facilidade. O ASQ difere do PSQ apenas em sua confiabilidade. Os alfas de Cronbach para o PSQ variaram entre 0,63 e 0,93 enquanto no ASQ todos os valores ultrapassaram 0,90. Sendo assim, pesquisadores de usabilidade devem utilizar o ASQ em vez do PSQ.

O PSSUQ é atualmente um instrumento de 19 questões para avaliar a satisfação do usuário com a usabilidade do sistema. Tendo em vista o número maior de questões, o participante necessita de mais tempo para concluir o PSSUQ do que o ASQ. A conclusão do PSSUQ fornece uma avaliação geral do sistema. O PSSUQ assim como o ASQ utiliza escalas gráficas de 7 pontos, ancoradas nos pontos finais com os termos "Concordo plenamente" em 1, e "Discordo plenamente" em 7. Os itens do PSSUQ avaliam características como: facilidade de uso, facilidade de aprendizado, simplicidade, eficácia, informações e

interface do usuário. As avaliações teste do PSSUQ são encontradas em [39] e [40]. Tais estudos forneceram evidências suficientes para usar o PSSUQ, mas também sugere que seria prudente coletar dados em circunstâncias diferentes para estender a generalização dos resultados.

O CSUQ é uma versão revisada do PSSUQ. O CSUQ é quase idêntico ao PSSUQ [41] exceto pelo fato da redação dos itens do questionário, que busca não explicitar que se trata de teste de usabilidade. Havia uma preocupação de que a configuração dos itens do PSSUQ pudessem ter influenciado nas correlações entre os itens e, consequentemente, nos fatores resultantes. Os resultados de uma avaliação teste do CSUQ evidenciaram que o questionário funciona bem em ambientes não laboratoriais. Em termos de confiabilidade e validade, os valores são comparados as análises do PSSUQ. Resumidamente, pesquisadores que realizam estudo de usabilidade podem utilizar o CSUQ para avaliação a satisfação do usuário em sistemas de usabilidade.

Outro questionário para avaliação de usabilidade é o System Usability Scale (SUS), desenvolvido por Brooke [8], contendo 10 questões que visam medir a usabilidade de um determinado produto. Em comparação com outros tipos de questionário, o SUS pode ser utilizado em uma gama maior de produtos: *websites*, *hardware*, sistemas de comando de voz, sistemas multimodais, entre outros. Os autores consideram que o SUS é um avaliador robusto e versátil que torna a pesquisa rápida e fácil. Outras vantagens do SUS seriam: (i) uso de uma escala de fácil aplicabilidade para os usuários; (ii) uso em tamanhos pequenos de amostras com resultados confiáveis; (iii) pode efetivamente diferenciar entre sistemas utilizáveis e inutilizáveis. O SUS produz um único número que representa uma medida composta da usabilidade geral do sistema a ser estudado. Os 3 pilares do SUS são: Efetividade, Eficiência e Satisfação. No questionário usado, as questões pares são questões negativas referentes a problemas do *software*, as questões ímpares são questões positivas referentes a qualidades do *software*. A escala Likert é utilizada neste questionário. O questionário de 10 questões é dividido da seguinte forma:

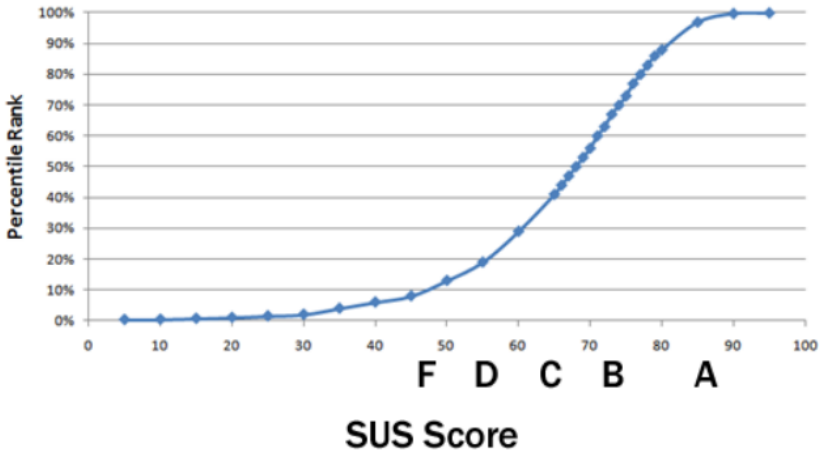
- Facilidade de aprendizagem: questões 3, 4, 7, 10;
- Eficiência: questões 5, 6, 8;
- Facilidade de memorização: questão 2;
- Minimização dos erros: questão 6;
- Satisfação: questões 1, 4, 9.

A contabilização do índice para avaliar a usabilidade pelo SUS é o resultado da fração entre a soma dos pontos das questões sobre o total de pontos possíveis. Para o cálculo, as questões ímpares, positivas, mantêm o valor avaliado. As questões pares, negativas, recebem o valor complementar. Por exemplo, se o participante der a nota 0 ("Discordo Totalmente"), escala de 0 a 4, na Questão 2, o resultado computado será 4.

Em um estudo com 500 avaliações [42] utilizando o SUS, encontrou-se uma curva de potencial, mostrada na Figura 7. Em média, as avaliações obtiveram resultado de 68% dos pontos possíveis com a utilização do questionário. Este percentual, pela curva de potencial, é denominado como sendo do *ranking* C. As outras classificações do *ranking* podem ser vistas na mesma figura. O eixo x (*SUS score*) representa o valor consolidado do resultado do questionário SUS. O eixo y (*Percentile Rank*) representa o percentil das 500 observações realizadas para construção da curva de potencial. Por exemplo, para

construção deste gráfico, um pouco mais de 10% das observações obteve 50 como o valor máximo consolidado do SUS. Por fim, as letra A, B, C, D e F representam, nesta ordem, o grau de qualidade de usabilidade do *software*. A é a melhor usabilidade e F, a pior. A curva de potencial é de grande utilidade para a avaliação da usabilidade de novos softwares, por fornecer uma base sólida para comparação.

Figura 7 – Curva de potencial - Score SUS.



Fonte: https:measuringu.com/sus/.

No Capítulo 2 são descritos os testes estatísticos utilizados neste trabalho, o projeto dos experimentos realizados e os testes para avaliação de aprendizagem e usabilidade escolhidos, bem como justificativas para tais escolhas. São também apresentados e descritos os questionários utilizados.

2 METODOLOGIA

Neste capítulo, apresentaremos a metodologia estatística utilizada para análise do estudo. Além disso, descreveremos cada detalhe do experimento realizado.

2.1 Testes Estatísticos

Nesta seção serão descritos os testes estatísticos utilizados neste trabalho.

Em ciências comportamentais busca-se chegar a uma decisão assertivamente em relação a uma hipótese realizando pesquisas, coletando dados empíricos e determinando o grau de aceitabilidade dessa hipótese em relação de conhecimentos e teorias do comportamento em estudo. Ou seja, deve-se verificar se evidências amostrais apoiam a hipótese testada, o que pode levar à confirmação, rejeição ou reformulação da hipótese. Além disso, é de suma importância dar ênfase à exigência científica de objetividade, visto que conclusões devem ser baseadas em métodos de conhecimento público e de uso geral. Observe-se que, mesmo com todos os cuidados, ainda há o risco de que a decisão considerada seja incorreta [43].

Para execução correta do experimento de avaliação, alguns aspectos devem ser seguidos e respeitados, para se obter um bom resultado ao final do experimento. São eles:

1. Definir a Hipótese Nula.
2. Escolher o método estatístico para realização do teste.
3. Determinar um nível α de significância.
4. Especificar ou inferir a distribuição amostral.
5. Definir a região de rejeição.
6. Calcular o valor da estatística teste e compará-lo à região de rejeição.

Cada um desses aspectos será melhor descrito a seguir.

2.1.1 Teste de Hipótese

É uma metodologia estatística que auxilia na tomada de decisões sobre uma ou mais populações baseado na informação obtida da amostra.

Na tomada decisões é conveniente a formulação de conjecturas sobre as populações de interesse no estudo, geralmente consistindo em afirmações acerca dos parâmetros (μ, σ^2, p) das mesmas, respectivamente média, variância e proporção. Essas suposições, que podem ou não ser verdadeiras, são denominadas *hipóteses estatísticas*.

Em muitas situações práticas o interesse do pesquisador é verificar a veracidade sobre um ou mais parâmetros populacionais (μ, σ^2, p) ou sobre a distribuição de uma variável aleatória.

Sendo assim, o Teste de Hipótese fornece informações que nos permitem rejeitar ou não rejeitar uma hipótese estatística através da evidência fornecida da amostra.

O Teste de Hipótese basicamente é formulado a partir de um ponto de partida, uma premissa, denominada *hipótese nula* (H_0). Ao conduzirmos um Teste de Hipótese, queremos evidenciar a rejeição da hipótese nula, ou seja, comprovar uma hipótese denominada de *hipótese alternativa* (H_1). Tais definições serão melhor apresentadas a seguir.

Hipótese Nula

Como processo introdutório para a tomada de decisão, necessita-se definir a hipótese nula (H_0), conjectura formulada com o propósito de rejeição, já que a intenção do estudo é provar ser verdadeira a hipótese alternativa (H_1), contrária a H_0.

> A hipótese alternativa é a definição operacional da hipótese de pesquisa do pesquisador. A hipótese de pesquisa é a predição deduzida da teoria que está sendo comprovada. Quando desejamos tomar uma decisão sobre diferenças, testamos H_0 contra H_1. H_1 é a asserção a ser aceita, se H_0 é rejeitada. [43].

No entanto, como pode-se presumir, um Teste de Hipótese não conduz necessariamente à decisão correta. Qualquer que seja a decisão tomada, há sempre a possibilidade de ocorrência de erros. Há 2 tipos de erros que podem-se cometer nos testes de hipóteses: erro do tipo I e erro do tipo II. As definições de erro são apresentadas na sequência.

Erros Decisórios para o Teste de Hipótese

Mesmo que todos os passos sejam respeitados para execução do experimento, dois erros podem acontecer, direcionando a afirmações incorretas a respeito das hipóteses formuladas.

- Erro do Tipo I

  Ocorre quando encontram-se evidências para rejeitar a hipótese nula que, no entanto, não deveria ser rejeitada. Denominamos *região crítica* (RC) como a *região correspondente à rejeição da hipótese nula*, e α como *nível de significância*, que é a probabilidade de ocorrência desse tipo de erro. Usualmente, dois valores para essa probabilidade são fixados como referência, 5% ou 1%. Sendo $\overline{x}$ a média encontrada na amostra do estudo para inferência acerca do parâmetro μ da população, tem-se que a probabilidade de $\overline{x}$ pertencer a RC (rejeitar H_0), dado que H_0 é verdadeira é descrita a seguir:

$$P(\text{erro do tipo I}) = P(\overline{x} \in RC \mid H_0 \text{ é verdadeira}) = \alpha$$

- Erro do Tipo II

  Ao não rejeitar a hipótese nula quando esta é falsa, comete-se o erro do tipo II. Sendo β a probabilidade de ocorrer o erro do tipo II, ou seja, a probabilidade de $\overline{x}$ não pertencer a RC dado que H_1 é verdadeira, tem-se que:

$$P(\text{erro do tipo II}) = P(\overline{x} \notin RC \mid H_1 \text{ é verdadeira}) = \beta$$

No entanto, a probabilidade β não pode ser aferida na maioria dos casos, o que leva, em muitas análises, a só considerar o erro do tipo I [44]. Além disso, apesar de α e β não possuírem relação exata, quanto menor for o valor de α, maior será o valor de β. A única forma de reduzir α e β simultaneamente é aumentando o tamanho da amostra.

Potência ou Poder do Teste

O *poder do teste* ou *potência* é a probabilidade de rejeitar a hipótese nula H_0, quando a hipótese alternativa H_1 é verdadeira. Ou seja, o Poder do Teste é o complemento do erro do tipo II, $1 - \beta$, conforme equação a seguir.

$$1 - \beta = P(\text{rejeitar } H_0 \mid H_1 \text{ verdadeira})$$

O poder do teste é afetado por 3 fatores:

- **Tamanho da Amostra**: Mantendo-se todos os outros parâmetros iguais, quanto maior o tamanho da amostra, maior o poder do teste;
- **Nível de Significância**: Quanto maior o nível de significância, maior o poder do teste. Se aumenta o nível de significância, reduz a região de aceitação (região em que não rejeita-se H_0). Como resultado, tem-se maior probabilidade de rejeitar a hipótese nula. Isto significa que tem-se menor probabilidade de aceitar a hipótese nula quando ela é falsa, isto é, menor probabilidade de cometer um erro do tipo II. Então, o poder do teste aumenta;
- **O verdadeiro parâmetro do valor a ser testado**: Quanto maior a diferença entre o verdadeiro valor do parâmetro e o valor especificado pela hipótese nula, maior o poder do teste.

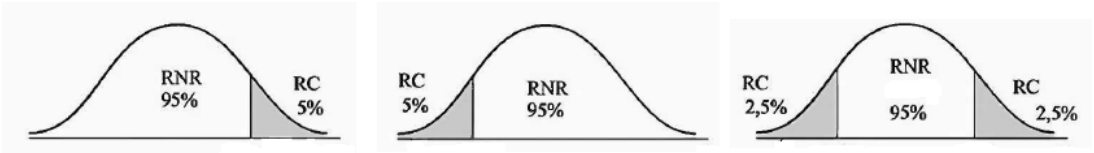
Teste Bilateral

Teste bilateral é aquele cujo valor α de significância é repartido de modo igual para as duas caudas da distribuição, representando a RC.

Neste caso, H_0 é a hipótese de igualdade e a H_1 é a hipótese da diferença. Vide Figura 8, sendo RNR denominada região de não rejeição. Sendo $\overline{x}$ o valor a ser testado e μ_0 o parâmetro populacional, as hipóteses formuladas tem a seguinte representação:

$$H_0 : \overline{x} = \mu_0$$
$$H_1 : \overline{x} \neq \mu_0$$

Figura 8 – 3 Tipos de Testes de Hipótese: (i)unilateral à direita; (ii) unilateral à esquerda; (iii)bilateral.



Fonte: (Morettin, 2010).

Teste Unilateral à Direita

Teste unilateral à direita é aquele cujo valor α de significância representa a cauda à direita da distribuição, demarcando a região crítica.

Neste caso, H_0 é a hipótese de igualdade e a H_1 é a hipótese em que sugere-se que o valor testado ($\overline{x}$) seja maior que o parâmetro populacional (μ_0). Vide Figura 8. As hipóteses formuladas tem a seguinte representação:

$$H_0 : \overline{x} = \mu_0$$
$$H_1 : \overline{x} > \mu_0$$

Teste Unilateral à Esquerda

Teste unilateral à esquerda é aquele cujo valor α de significância representa a cauda à esquerda da distribuição, demarcando a região crítica.

Neste caso, H_0 é a hipótese de igualdade e a H_1 é a hipótese em que sugere-se que o valor testado ($\overline{x}$) seja menor que o parâmetro populacional (μ). Vide Figura 8. As hipóteses formuladas tem a seguinte representação:

$$H_0 : \overline{x} = \mu_0$$
$$H_1 : \overline{x} < \mu_0$$

Sendo assim, deve-se escolher qual teste mais adequa-se a determinado estudo. Em seguida, utilizar um dos 3 métodos para testar a hipótese e verificar se deve-se ou não rejeitar H_0. Esses 3 métodos são:

- Método do Intervalo de Confiança;
- Método da Região Crítica;
- Método do p-valor.

Todos os 3 testes são equivalentes para análise. Neste caso, para este estudo, optou-se por utilizar nas análises com teste de hipótese o método do p-valor que será descrito a seguir.

Método do p-valor

O *p-valor* de um teste é o menor valor de significância que nos leva a rejeitar H_0. Também pode ser denominado como *nível descritivo* ou *probabilidade de significância.* O p-valor permite testar hipóteses de forma direta, sem a necessidade do uso de tabelas. A regra da decisão é a seguinte: se p-valor $\leq \alpha$, rejeitamos H_0. Caso contrário, não rejeitamos.

O p-valor de um teste é dado pela probabilidade, calculada sob H_0, de que a estatística do teste assuma um valor igual ou mais extremo do que o valor calculado na amostra. Neste caso, se era pouco provável que ocorre o valor de teste e mesmo assim ocorreu, devemos rejeitar H_0.

O próximo passo é escolher o método estatístico para realização do teste de hipótese. Para isso necessita-se definir o modelo estatístico da distribuição dos dados. A distribuição dos dados é um dos pressupostos necessários para definição de qual técnica será utilizada para avaliar as hipóteses de um teste. Sendo assim, define-se a seguir pressupostos e argumentos para definição do modelo estatístico dos dados.

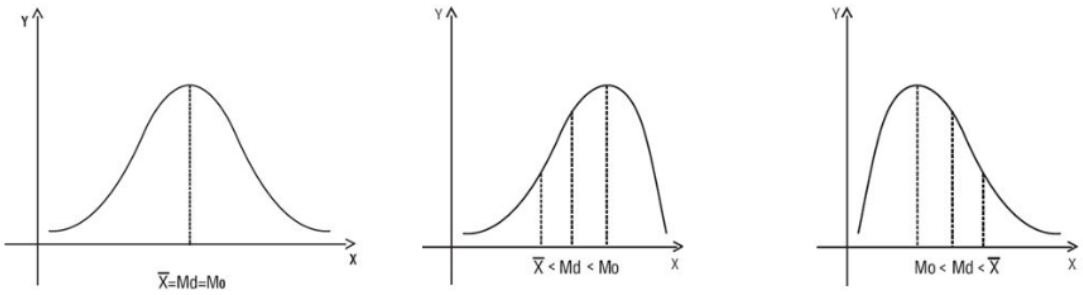
2.1.2 Definição do Modelo Estatístico

O Modelo Estatístico é estabelecido ao determinar a natureza da população e o modo de amostragem utilizado diante das condições de mensuração. Em alguns casos, apura-se que as condições de determinado modelo são satisfeitas. No entanto, na grande maioria, deve-se apenas admitir que satisfaçam. Sendo assim, são necessárias suposições para definição da estatística teste a ser utilizada. Provas que comportam maior poder de avaliação (poder do teste) tais como os Testes Paramétricos t e F, necessitam de suposições mais fortes. Por exemplo, para utilização do teste t ou teste F, tem-se, obrigatoriamente, que cumprir no mínimo as condições a seguir:

- As observações devem ser independentes.
- Distribuição normal.
- Mesma variância entre as populações.
- Variável em escala intervalar, no mínimo, ou seja, não pode estar na escala nominal ou ordinal. Escalas nominais e ordinais são escalas que representam categorias qualitativas. A diferença entre as duas é que, na escala ordinal as categorias são ordenadas.
- As médias populacionais devem ser combinações lineares. Esta condição aplica-se somente ao teste F.

No entanto, nem sempre as condições anteriores são satisfeitas. Com isso, o emprego de testes não-paramétricos torna-se coerente e importante. Além disso, a maior parte dos testes não-paramétricos são aplicados em dados ordinais e, em alguns casos, em escalas nominais.

Figura 9 – 3 Tipos de Distribuições: Distribuição simétrica; Assimétrica à esquerda; Assimétrica à direita.



Fonte: (Morettin, 2010).

A Figura 9 representa 3 tipos de distribuições. A distribuição normal e duas distribuições assimétricas, à esquerda e à direita. Na distribuição normal, a média ($\overline{x}$), a moda (Mo) e a mediana (Md) estão representadas no mesmo ponto do gráfico, sendo uma boa representação da massa de dados. No entanto, nas duas distribuições assimétricas a mediana representa uma medida de posição mais representativa que a média. Os testes não-paramétricos realizam o teste de hipótese através da mediana, enquanto os testes paramétricos analisam a média. Esse é um dos motivos pelos testes não-paramétricos serem mais utilizados em dados com distribuição assimétrica.

Torna-se importante então, avaliar se a distribuição dos dados em teste segue uma distribuição normal e, para isso, há alguns testes de hipótese importante além da utilização de técnica gráfica. A técnica gráfica e um dos testes de normalidade serão apresentados a seguir.

Teste de Normalidade para Distribuição dos Dados

Para realização de testes estatísticos necessita-se conhecer a distribuição dos dados em estudo, mais precisamente, identificar se os dados seguem ou não uma distribuição Normal. Os testes de normalidade são utilizados para verificar se a distribuição associada a um conjunto de dados pode ser aproximada pela normal. As principais técnicas encontradas na literatura são: técnica gráfica, teste de *Kolmogorov-Smirnov*, teste de *Anderson-Darling*, teste de *Shapiro-Wilk* e teste de *Ryan-Joiner*.

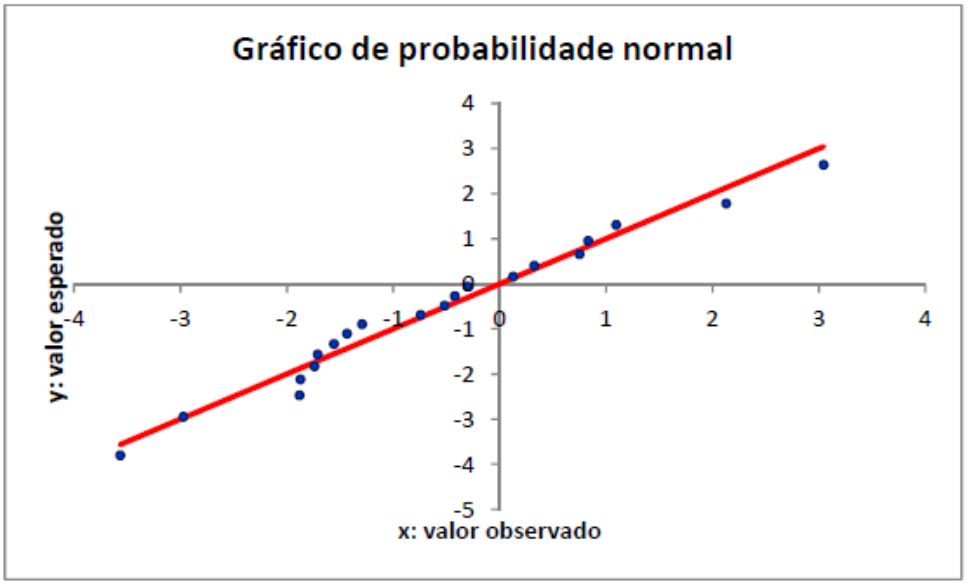
Para este estudo, optou-se por utilizar a técnica gráfica e observar através dela se há ou não normalidade da distribuição dos dados. Tal técnica possibilita avaliar se há, pelo menos, simetria na distribuição. Além disso, optou-se também por utilizar a estatística do Teste de *Shapiko-Wilk* por possuir maior rigidez em sua avaliação [45].

Técnica Gráfica

A técnica gráfica é realizada através de um gráfico de probabilidade com o objetivo de verificar se um conjunto de dados pode ter sido gerado a partir de uma específica distribuição de probabilidade contínua.

A lógica da construção deste gráfico é a comparação entre os dados observados na amostra (x) com os dados que esperaríamos ter observado caso eles seguissem uma determinada distribuição de probabilidade, no caso, distribuição normal. Caso fosse possível criar uma coluna (y) com os dados esperados e dispuséssemos os pontos (x,y) em um eixo cartesiano esperaríamos que, se os dados seguem de fato a distribuição proposta, os pontos se distribuíssem aleatoriamente ao redor da reta. Vide Figura 10.

Figura 10 – Exemplo de um gráfico de probabilidade para distribuição Normal.



Fonte: http://www.portalaction.com.br/.

Seja uma amostra de tamanho n com média $\overline{x}$, desvio-padrão s e x_i a i-ésima observação. Seja $\hat{F}(i)$ a função que retorna percentis, espaçados de forma homogênea, de acordo com o tamanho da amostra, tem-se:

$$\hat{F}(i) = \frac{i - 0{,}5}{n}$$

Sendo assim, utiliza-se a função inversa da distribuição de interesse. Se a distribuição de interesse for a distribuição normal tem-se: $\hat{K}^{-1}_{normal}$ a função inversa da distribuição normal que retorna o dado esperado (y_i) da i-ésima observação tem-se:

$$y_i = \hat{K}^{-1}_{normal}\left(\hat{F}(i), \overline{x}, s\right)$$

O próximo passo é dispor os pares (x,y) no eixo cartesiano e comparar a disposição dos pontos com a reta esperada da distribuição. A reta esperada é uma reta com ângulo de 45º, onde, o valor esperado é igual ao valor observado. Se os valores se aderem bem a reta, podemos considerar que seguem a distribuição proposta. Em caso contrário, os valores seguem outro tipo de distribuição.

Esta técnica pode ser utilizada para verificar a aderência de uma determinado conjunto de dados a qualquer distribuição de probabilidade. Basta, para isso, utilizar a função de distribuição acumulada correspondente.

Teste de Shapiro-Wilk

O teste de *Shapiro-Wilk*, proposto por Shapiro [46], é baseado na estatística W dada por:

$$W = \frac{b^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Neste caso, x_i são os valores da amostra ordenados, $\overline{x}$ é a média amostral e b é uma constante determinada da seguinte forma, se n (tamanho da amostra) é par, o cálculo de b é dado por:

$$b = \sum_{i=1}^{\frac{n}{2}} a_{n-i+1}(x_{n-i+1} - x_i)$$

Em caso contrário, ou seja, se n é ímpar, o cálculo de b é dado da seguinte forma:

$$b = \sum_{i=1}^{\frac{n+1}{2}} a_{n-i+1}(x_{n-i+1} - x_i)$$

Em que a_{n-i+1} são constantes geradas pelas médias, variâncias e covariâncias das estatísticas de ordem de uma amostra de tamanho n de uma distribuição Normal. Seus valores, tabelados, são encontrados no Apêndice 3.

Define-se então os seguintes passos:

1. Formulação da Hipótese;
    - Hipótese Nula: A amostra provém de uma população Normal;
    - Hipótese Alternativa: A amostra não provém de uma população Normal.

2. Estabelecer o nível de significância α. Geralmente igual à 0,05;

3. Calcular a estatística teste W;

4. Tomada de decisão. Se p-valor $< \alpha$ rejeita-se a Hipótese Nula.

Sendo assim, determinada a normalidade ou não da distribuição, opta-se pelo teste adequado para avaliação desejada e correta dos dados. Na próxima seção, apresenta-se o testes paramétricos e não paramétricos utilizados no estudo.

2.1.3 Teste Paramétrico - Teste t pareado

Para realizar um teste de igualdade entre médias e variâncias necessita-se que duas populações sejam independentes. No entanto, na prática, geralmente encontram-se populações dependentes. Neste caso, utiliza-se o teste t pareado.

Sejam duas amostras dependentes $X_1, X_2, ..., X_n$ e $Y_1, Y_2, ..., Y_n$, pareadas da forma $(X_1,Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ e $D_i = X_i - Y_i$ para $i = 1, 2, 3, ..., n$. Apesar das amostras serem dependentes considera-se que D_i tenha distribuição normal:

$$D_i \sim N(\mu_D,\sigma_D^2)$$

O parâmetro μ_D é estimado por $\overline{D}$, média amostral das diferenças. Já o parâmetro σ_D^2 é estimado pela variância amostral das diferenças, ou seja:

$$s_D^2 = \frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1}$$

Por fim, a estatística t será calculada pela seguinte expressão:

$$T = \frac{\overline{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}$$

A distribuição é uma T de Student com $n-1$ graus de liberdade. Considerando α o nível de significância, tem-se os pontos críticos $t_{\frac{\alpha}{2}}$ e $-t_{\frac{\alpha}{2}}$, mostrados na Figura 11.

Para definir a rejeição de H_0 necessita que $T > t_{\frac{\alpha}{2}}$ ou $T < -t_{\frac{\alpha}{2}}$.

Outra possibilidade para rejeição de H_0 é com o cálculo do *p-valor*, definido como:

$$\text{p-valor} = P[|t| > |T||H_o] = 2P[|t| > |T||H_o]$$

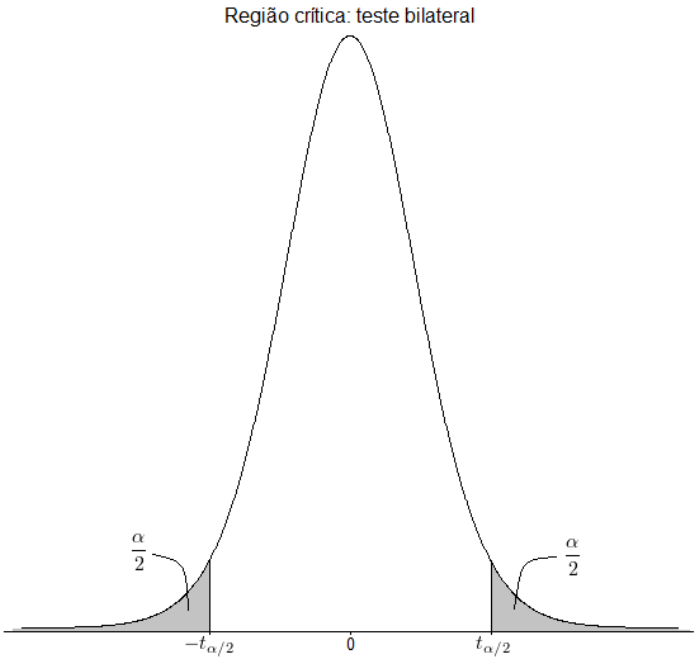
Se p-valor $< \alpha$ rejeita-se a hipótese nula. Com isso então, define-se uma estatística paramétrica para análise de distribuições com amostras dependentes pareadas.

A seguir, definem-se algumas ferramentas de análise não-paramétricas, suas vantagens e desvantagens em relação à estatística paramétrica, além de um teste correlato ao teste t.

2.1.4 Vantagens e Desvantagens da Estatística Não-Paramétrica

Para a utilização de qualquer técnica estatística há a necessidade de conhecer suas vantagens e desvantagens. No caso dos testes Não-Paramétricos as afirmações probabilísticas, em geral são exatas e independem da distribuição populacional. Ademais, não há opção estatística para testes em amostras com tamanho muito pequeno nem para tratamento de amostras de populações diferentes, o que torna os testes não-paramétricos a melhor opção. Outras vantagens dos testes não-paramétricos são que eles permitem o

Figura 11 – Regiões críticas e pontos críticos - Teste t pareado.



Fonte: http://www.portalaction.com.br/.

tratamento de dados nominais, ou seja, dados qualitativos, além da fácil aplicabilidade e aprendizado.

Por outro lado, se as condições para utilização de testes paramétricos são satisfeitas e todos os pressupostos são atendidos, utilizar provas não-paramétricas consiste em um desperdício de informação, pois os testes paramétricos tem maior poder de decisão. Também, não há testes não-paramétricos para testar interações em modelos de análise de variância.

Poder-Eficiência

Sabe-se que a força das asserções feitas para a escolha do teste está diretamente relacionada com a generalização das conclusões resultantes, ou seja, quanto maior a generalização menos poderosa será a prova de H_0. Resumidamente, para tamanhos de amostras iguais, um teste que necessita de mensurações mais fracas terá, normalmente, menos poder de decisão. No entanto, para tamanhos diferentes, isto pode não ser verdadeiro. Por exemplo, um teste X pode ser mais fraco que um teste Y para uma amostra de tamanho $N = 50$. No entanto, X poderá ser mais forte com um tamanho amostral $N = 40$, comparado à prova Y com tamanho $N = 30$.

"O conceito de poder-eficiência está ligado ao aumento do tamanho da amostra necessária para tornar a prova B tão poderosa quanto a prova A" [43].

Considerando A a prova mais poderosa para um estudo, B outra ferramenta pertinente e que B é tão poderosa com tamanho amostral N_b quanto A com tamanho N_a tem-se:

$$\text{Poder-Eficiência do teste B} = (100)\frac{N_a}{N_b}\%$$

2.1.5 Testes Não-Paramétricos para duas amostras

Os testes estatísticos para duas amostras são empregados no momento em que o experimento é estabelecido com dois tratamentos diferentes. Ou seja, quer se comparar um grupo que teve tratamento com o grupo que não teve ou recebeu tratamento diferente.

> Por exemplo, um pesquisador pode tentar comparar dois métodos teóricos, tendo um grupo de alunos ensinado por um método e um grupo diferente ensinado por outro. Agora, se um dos grupos tem alunos mais capazes ou mais motivados, o desempenho dos dois grupos após as diferentes experiências de aprendizado pode não corresponder de forma precisa à eficácia relativa dos dois métodos de ensino, porque outras variáveis estão criando diferenças no desempenho.[47].

Teste de Wilcoxon pareado

O Teste de Wilcoxon é similar ao Teste dos Sinais e é o correspondente não-paramétrico do teste t pareado. Enquanto, o Teste dos Sinais avalia a diferença entre as condições antes e após o tratamento, o Teste de Wilcoxon pareado considera também o valor dessas diferenças, ou seja, maior ponderação a um par que possua uma grande diferença entre as condições, seja positiva ou negativa. Consequentemente será uma ferramenta mais poderosa de análise.

Para aplicação do teste de Wilcoxon, seja d_i o módulo da diferença entre pares. Atribui-se a eles postos, do menor ao maior d_i. Posteriormente, indica-se quais destes postos são oriundos de diferenças negativas ou positivas. Caso algum par possua diferença igual a 0, deve-se excluí-lo da análise e, consequentemente, reduzir o tamanho N da amostra em uma unidade. Por outro lado, se houver pares com diferenças iguais, o número do posto recebido por eles será igual, sendo calculado pela média aritmética do valor que teriam recebido caso fossem diferentes.

Resumindo, o procedimento segue os seguintes passos:

1. Calcular o módulo e indicar o sinal da diferença de cada par.
2. Atribuir os postos.
3. Indicar a natureza de cada posto, positiva ou negativa.
4. Determinar T, como a menor soma de postos de mesmo sinal.
5. Determinar N, como sendo o número de postos utilizados em T.
6. Calcular o valor crítico e compará-lo ao observado.
   (a) Se $N \leq 25$, conferir o valor crítico no Apêndice 3.
   (b) Se $N > 25$, calcular o valor de z através da aproximação da normal, encontrar o valor p e compará-lo ao α.
7. Por fim, se o valor de p encontrado for menor que α, rejeita-se H_0.

Quando $N > 25$, a soma dos postos tem distribuição normal, com média e desvio padrão:

$$\text{Média} = \mu_T = \frac{N(N+1)}{4}$$

$$\text{Desvio-Padrão} = \sigma T = \sqrt{\frac{N(N+1)(2N+1)}{24}}$$

Logo, z padronizado será:

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}$$

Poder-Eficiência

Quando as suposições da prova paramétrica t são realmente satisfeitas, a eficiência assintótica, na vizinhança de H_0, da prova de Wilcoxon comparada com a prova t é $3/\pi = 95{,}5\%$. Significa isto que $3/\pi$ é o limite da razão dos tamanhos de amostras necessários para que tanto a prova de Wilcoxon como a prova t tenham o mesmo poder. Para pequenas amostras, a eficiência é próxima de 95%. [43].

Teste U de Mann-Whitney

Torna-se difícil realizar estudos com amostras relacionadas, principalmente porque um indivíduo só pode realizar uma determinada tarefa não familiar uma única vez, e também quando a natureza da variável possa impedir o uso do indivíduo como seu próprio controle. Nesses casos aplicam-se amostras independentes, que podem ser extraídas aleatoriamente de populações diferentes sem exigir que as amostras possuam o mesmo tamanho.

No entanto, pelo exposto acima, ferramentas como o Teste dos Sinais e Wilcoxon Pareado não podem ser aplicados para este tipo de amostragem. Contudo, há ferramentas para análise de amostras independentes, como por exemplo, o Teste U de Mann-Whitney.

O teste U de Mann-Whitney é uma das provas não-paramétricas mais poderosas, exigindo grau mínimo de mensuração ordinal e inferior à escala intervalar.

Para sua aplicação, sejam n_1 e n_2, respectivamente, o número de casos no menor e no maior dos dois grupos independentes. Inicialmente, juntam-se as observações ou escores dos dois grupos em ordem crescente. Desta vez, considera-se o valor algébrico e atribui-se, também em ordem crescente, os postos para cada observação. Em caso de empate entre escores, realiza-se o mesmo processo do teste Wilcoxon pareado.

Observando então o grupo 1, o valor da estatística U é dado pelo número de vezes que uma observação no grupo 2 precede um escore do grupo 1 em classificação crescente.

Resumidamente, procedimento segue os passos:

1. Determinar n_1 e n_2.
2. Unir as observações dos dois grupos em ordem crescente, classificando cada um com um posto.
3. Determinar o valor de U.

4. Encontrar o valor crítico para a Estatística U.

   (a) Se $n_2 \leq 8$, verificar a Tabela de Mann-Whitney para valores menores ou iguais a 8.

   (b) Se $9 \leq n_2 \leq 20$ verificar a Tabela de Mann-Whitney no Apêndice 3 para valores entre 9 e 20.

   (c) Se $n_2 > 20$ aplicar a aproximação da normal, calcular z e comparar o resultado com o valor p. Aplicar a correção de empates caso a proporção destes seja muito grande.

5. Por fim, rejeitar H_0 se o valor observado de Mann-Whitney for inferior a α.

Para determinar o valor de U usamos uma das duas expressões equivalentes abaixo:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

ou,

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Já para grandes amostras, a distribuição de U rapidamente tende à distribuição normal. Logo:

$$\text{Média} = \mu_U = \frac{n_1 n_2}{2}$$

e

$$\text{Desvio-Padrão} = \sigma_U = \sqrt{\frac{(n_1)(n_2)(n_1+n_2+1)}{12}}$$

Pode-se determinar a significância de U observado com distribuição normal com média 0 e variância 1 pela expressão:

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{(n_1)(n_2)(n_1+n_2+1)}{12}}}$$

Poder-Eficiência

Calculando o poder eficiência do teste, frente a um conjunto de dados em que possa também aplicar a prova t, o poder-eficiência de U de Mann-Whitney tenderá à $\frac{3}{\pi} \approx 95{,}5\%$ para amostras consideradas grandes. Em outros casos, se aproxima de 95%.

A seguir apresenta-se um coeficiente que avalia a consistência interna de questionários medem fatores subjetivos.

2.1.6 Alfa de Cronbach

Uma maneira de medir fatores subjetivos é questionar o que pensamos estar relacionado com o fator, dando a cada resposta uma pontuação e somarmos os scores obtidos. Quando itens são usados para a construção de uma escala eles precisam ter consistência interna, ou seja, têm que medir o mesmo tipo de fenômeno. O Coeficiente α de *Cronbach* mede a consistência interna.

O cálculo de α é baseado na correlação média entre pares de itens que formam a escala. O α terá sempre um valor entre 0 e 1. Se houver perfeita correlação, então α é igual a

1. Caso os itens sejam completamente não correlacionados, então α é igual a 0. Se o α encontrado for alto, isto indica uma alta consistência interna.

Para identificar uma consistência interna satisfatória é necessário encontrar um valor de $\alpha \geq 0{,}7$. No entanto, um valor muito alto para α pode indicar simplesmente que há excesso de perguntas envolvidas na construção da escala no questionário, ou seja, alguns itens são desnecessários.

O α de *Cronbach* é aplicado para grupos de itens que, supostamente, medem diferentes aspectos de um mesmo conceito e indica quão bem diferentes itens os medem. O α de *Cronbach* reflete a homogeneidade de itens diferentes que medem a mesma variável ou qualidade.

Idealmente, as medidas representadas com escalas teriam que indicar sempre os mesmos resultados. Entretanto, no mundo real há uma série de fatores externos (aleatórios) que podem afetar a maneira com que os entrevistados respondem as perguntas. Busca-se, então, uma medida que quantifique isto, composta de 2 fatores: um teórico "Valor Verdadeiro" da escala e "a variação causada por fatores aleatórios". Esta relação é sumarizada na equação:

$$M = T + \xi,$$

em que M é a medida ou escala em questão, T é o valor verdadeiro teórico e ξ é o valor aleatório. O ξ pode ser tanto positivo quanto negativo. O valor verdadeiro T representa a média de pontuação que seria obtida se uma pessoa respondesse à questão infinitas vezes.

Sendo assim o α de *Cronbach* mede a correlação das respostas em um questionário através da avaliação do perfil das respostas dos entrevistados. O α é calculado a partir do somatório da variância dos itens e da soma da variância dos avaliadores pela equação:

$$\alpha = (\frac{k}{k-1}) * (1 - \frac{\sum_{i=1}^{k} s_i^2}{s_t^2}),$$

em que k corresponde ao número de itens do questionário, s_i^2 representa a variância de cada item e s_t^2 representa variância total do questionário (soma das variâncias dos avaliadores).

2.2 Experimento

Nesta seção, descreveremos o experimento conduzido no estudo deste trabalho.

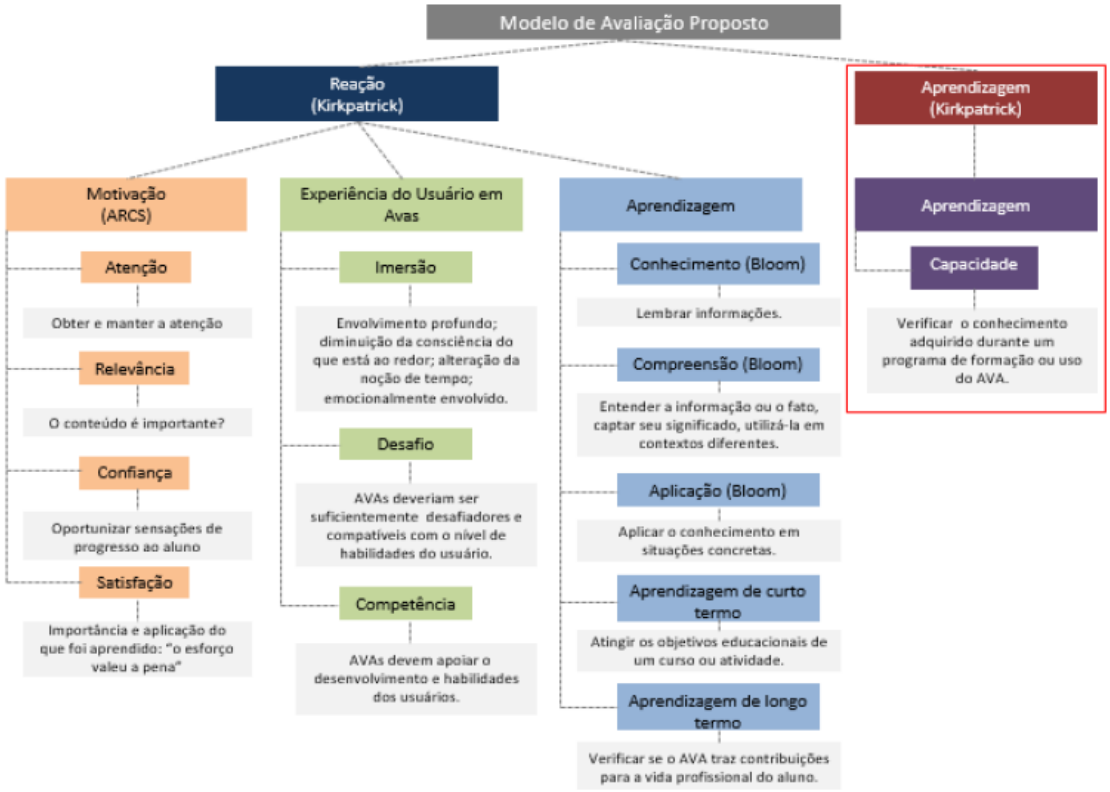
Considerando os estudos apresentados no capítulo anterior, optou-se por adotar as seguintes bases metodológicas para a avaliação de aprendizagem com o TuPy Online: uma adaptação da metodologia de Savi [3] com a contribuição de Ferreira [25] e a inclusão de grupos de controle, conforme detalhado a seguir.

Em relação a usabilidade, considerando também os estudos do capítulo anterior, foi escolhido o questionário SUS. Apesar de todos os questionários apresentados atenderem o propósito para análise da usabilidade do TuPy Online, a escolha do SUS foi devida ao menor número de questões que emprega e por sua qualidade em verificar usabilidade ser equivalente aos demais. Outro ponto positivo do SUS é que a inclusão de perguntas positivas e negativas de maneira alternada, facilita a identificação de respostas incoerentes, o que permite a remoção de tais respostas da análise final.

A Figura 12 resume todos os fatores que consideramos para elaboração do principal questionário da avaliação do TuPy Online. Essencialmente, este modelo tem a mesma base metodológica presente em Ferreira [25]. Os 3 grupos de fatores pertencentes ao nível 1

(Reação) são originários do modelo proposto por Savi et al. e estão detalhados à esquerda da figura. Estes fatores são abordados no principal questionário utilizado descrito na sequência. À direita, a contribuição de Ferreira com a inclusão do nível 2 (Aprendizagem) de Kirkpatrick. Tal contribuição corresponde à aplicação de um pré-teste e um pós-teste, conforme melhor descrito adiante.

Figura 12 – Estrutura do Modelo de Avaliação do MA-AVA.



Fonte: (Ferreira, 2017).

Savi et al. consideraram Reação em 3 fatores a serem avaliados: Motivação, Experiência e Aprendizagem. O fator Motivação foi descrito pelo modelo *ARCS de Keller*. O fator Aprendizagem foi descrito pela Taxonomia de Bloom e pela metodologia de [16]. O Fator Experiência foi descrito por 5 fatores: Imersão, Diversão, Controle, Desafio e Competência.

Passamos a descrever o questionário aplicado. Da mesma forma que em [25] fizemos uma adaptação própria do questionário de Savi et al.. Retiramos as questões consideradas não aplicáveis para um *software* educacional e incluímos toda a categoria Imersão. Diferentemente do questionário de Ferreira, mantivemos a categoria Controle no fator Experiência, pela importância na avaliação do TuPy Online. A parte principal do questionário resultante é apresentada na Tabela 6. Esta parte do questionário foi elaborada com 19 questões considerando as dimensões: Atenção, Relevância, Confiança, Satisfação, Desafio, Diversão, Competência e Controle. Além disso, consideramos também a Aprendizagem de curto termo e a Aprendizagem de longo termo.

Tabela 6 – Questionário para avaliação de percepção do TuPy.

QUESTIONÁRIO PARA AVALIAÇÃO DE PERCEPÇÃO DO TuPy	
Atenção	O design do TuPy é atraente
	O TuPy me pareceu interessante desde a apresentação inicial
	A variação (forma ou de atividades) ajudou a me manter atento ao TuPy
Relevância	O conteúdo do TuPy é relevante aos meus interesses
	O funcionamento do TuPy está adequado para o aprendizado
	O conteúdo do TuPy está conectado com outros conhecimentos que já possuia
Confiança	Foi fácil entender o TuPy e começar a utilizá-lo como material de estudo
Satisfação	Estou satisfeito porque sei que terei oportunidades de utilizar na prática coisas que aprendi com o TuPy
Desafio	O TuPy é adequadamente desafiador para mim, não é nem muito fácil nem muito difícil
Diversão	Diverti-me com o TuPy
	Eu recomendaria o TuPy para meus colegas
	Voltei a utilizar o TuPy outras vezes
Competência	Consegui atingir os objetivos do TuPy por meio de minhas habilidades
	Tive sentimentos positivos de eficiência no aprendizado de algoritmos
Controle	Os comandos para realizar ações no TuPy responderam bem
	É fácil aprender a usar a interface e comandos do TuPy
Aprendizagem de curto termo	O TuPy contribuiu para a minha aprendizagem na disciplina
	O TuPy foi eficiente para minha aprendizagem em comparação com outras atividades da disciplina
Aprendizagem de longo termo	A experiência com o TuPy vai contribuir para meu desempenho na vida profissional

Fonte: O próprio autor.

Anexo ao questionário de percepção, há uma seção de auto-avaliação, seguindo os princípios da Taxonomia de Bloom, segundo Tabela 7. O estudante indicaria um valor entre 1 e 4 (Fraco, Regular, Bom, Ótimo) para seu conhecimento anterior e posterior à utilização do TuPy Online, dividido em 3 categorias (Lembrança, Compreensão e Aplicabilidade) para diversos assuntos. Cada um desses assuntos refere-se à tópicos das disciplinas das turmas que serão avaliadas.

Tabela 7 – Taxonomia de Bloom utilizada por Savi et al.

AED1	Lembrança		Compreensão		Aplicabilidade	
	Antes	Depois	Antes	Depois	Antes	Depois
Assunto 1						
Assunto 2						
Assunto 3						
1 - Fraco 2 - Regular 3 - Bom 4 - Ótimo						

Fonte: O próprio autor.

O questionário para avaliação da usabilidade (SUS) foi utilizado em sua íntegra, como mostra a Figura 13. As 10 questões do SUS são avaliadas da seguinte forma: a satisfação do usuário; a facilidade de memorização; as inconsistências; a eficiência; e a facilidade de aprendizagem.

Figura 13 – SUS - Aplicado para avaliação do TuPy Online.

Eu acredito que gostaria de usar esse sistema com frequência	0 1 2 3 4 NA
Eu acho o sistema desnecessariamente complexo	0 1 2 3 4 NA
Eu achei o sistema fácil de usar	0 1 2 3 4 NA
Eu acho que precisaria de ajuda técnica para ser capaz de usar esse sistema	0 1 2 3 4 NA
Eu achei várias funções do sistema bem integradas	0 1 2 3 4 NA
Eu acho que há muitas inconsistências nesse sistema	0 1 2 3 4 NA
Eu acredito que muitas pessoas podem aprender a usar esse sistema facilmente	0 1 2 3 4 NA
Eu achei o sistema muito complicado de usar	0 1 2 3 4 NA
Eu me senti muito confiante usando o sistema	0 1 2 3 4 NA
Eu preciso aprender muitas coisas antes de poder usar esse sistema	0 1 2 3 4 NA

Fonte: O próprio autor.

Além disso, foi proposto, junto ao questionário SUS, um conjunto de perguntas buscando um *feedback* específico sobre possíveis melhorias do *software*. Junto à essas perguntas também foram propostos campos para o aluno expressar sua opinião em relação aos aspectos da linguagem, interface, e processo de introdução ao TuPy Online. Por fim, o aluno assinalou quais algoritmos da disciplina cursada ele simulou em casa. Vide Apêndice 3.

Em seguida, o experimento será descrito detalhadamente.

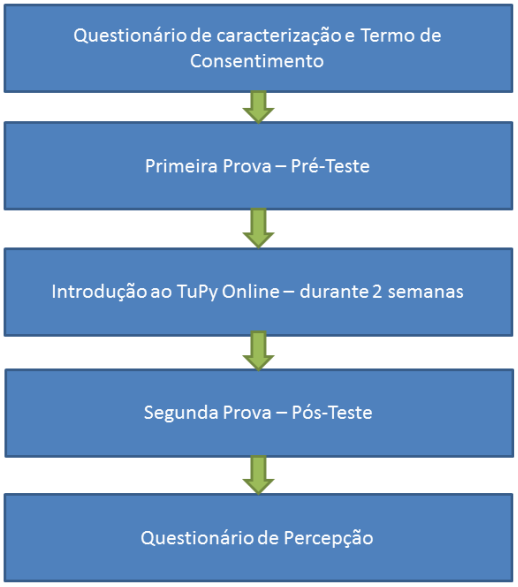
O experimento foi realizado na Universidade do Estado do Rio de Janeiro (UERJ), com alunos da graduação em Ciência da Computação, em turmas de disciplinas e professores diferentes, sendo: uma turma de Otimização em Grafos (OTG), duas turmas de Algoritmos e Estruturas de Dados I (AED1) e uma turma de Algoritmos e Estruturas de Dados II (AED2). Além disso, foi também aplicado a uma turma no curso de mestrado da disciplina obrigatória de Algoritmos (ALG).

O experimento foi projetado da seguinte forma: os alunos fariam em torno de um mês de aula com a metodologia tradicional de ensino utilizada na universidade e, em seguida, seriam informados sobre o experimento. Os alunos assinariam um termo de consentimento e utilização anônima de seus dados e preencheriam um questionário de caracterização de perfil, apresentados nos Apêndice 3 e 3. As principais informações do questionário de caracterização são o tempo disponível para utilização da ferramenta em casa, o desempenho do aluno na universidade e o conhecimento prévio de ferramentas que auxiliam no aprendizado de computação. Na aula posterior ao preenchimento deste questionário preliminar, realizariam a primeira prova, considerada, para efeitos do experimento, como um pré-teste. Após a primeira prova, durante um período de duas semanas, os estudantes seriam apresentados ao TuPy Online e revisariam todo o conteúdo do primeiro mês de aula por meio de exercícios, demonstrações e visualizações personalizadas desta ferramenta. Ao término deste período, nova prova seria aplicada para efeito de pós-teste. Na aula seguinte, os alunos responderiam os questionários de percepção, SUS e de *feedback*, apresentados no Apêndice 3 e Apêndice 3 e descritos detalhadamente anteriormente. Os alunos teriam a opção de participar ou não do experimento.

Foi adotada a ideia de grupo de controle para avaliar o aprendizado, mantendo o foco

no uso da ferramenta, e visando excluir outras variáveis tais como: modelos de provas distintas, qualidade de ensino intrínseco a cada professor, alunos com pré-conhecimento do conteúdo da disciplina. O experimento é esboçado graficamente na Figura 14.

Figura 14 – Projeto do experimento.



Fonte: O próprio autor.

O grupo de controle foi formulado da seguinte forma: as provas do pré-teste serão divididas em dois modelos A e B e cada turma dividida em dois grupos de mesmo tamanho. O grupo que fez no pré-teste a prova A, realizou a prova B no pós-teste. Já o outro, procedeu inverso, isto é, foi aplicado à prova B no pré-teste seguida da prova A no pós-teste.

A análise dos dados foi feita separadamente por turma e também em conjunto. Foram utilizados os testes, t pareado, Wilcoxon pareado e U de Mann-Whitney de acordo com o tamanho das turmas e grupos utilizados. Os resultados serão apresentados com o nível de confiança de 95%.

3 APRESENTAÇÃO DOS RESULTADOS

Nesta seção, apresentam-se os dados, resultados, considerações, análises e discussões do experimento realizado.

Em linhas gerais o experimento ocorreu conforme previsto e despertou um grande interesse inicial dos alunos. Obteve-se participação voluntária de todos os alunos das turmas selecionadas. Na primeira aula, realizou-se uma apresentação introdutória do TuPy Online pelos criadores da ferramenta. Em seguida apresentou-se aos alunos a execução de alguns exemplos de algoritmos estudados pela metodologia tradicional e anterior ao pré-teste. Nas outras aulas, os alunos tiveram a oportunidade de compilar seus próprios algoritmos. Além disso, alguns exercícios para casa foram propostos a fim de aumentar o tempo em contato com o *software*. Os alunos, por fim, relataram o tempo de utilização do TuPy Online em casa.

Um importante ponto do estudo que deve ser mencionado é que, como a participação era facultativa, alguns alunos que realizaram o pré-teste e pós-teste não responderam o questionário final, bem como alguns que, por algum motivo, não participaram de um dos testes, quiseram responder os questionários e foram considerados válidos.

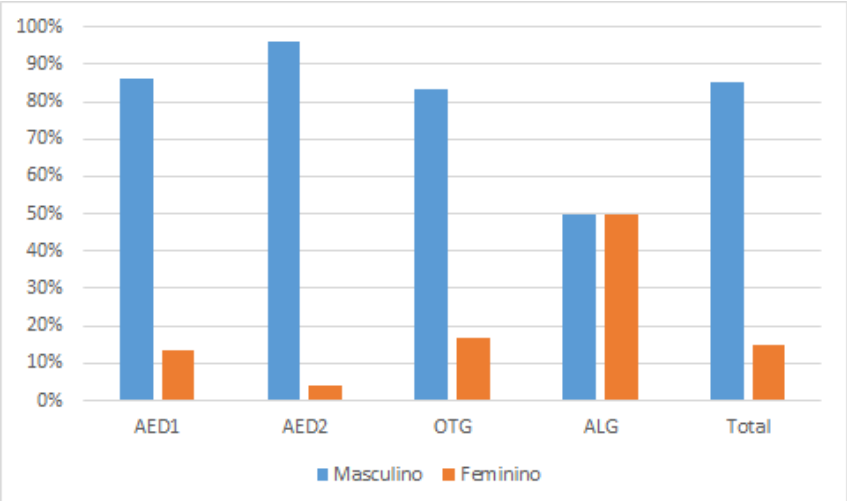
Outro ponto importante é que uma das turmas de AED1 tinha poucos alunos efetivamente acompanhando a disciplina. Consequentemente, esta turma foi retirada das avaliações a fim do não enviesamento dos resultados.

Cabe ressaltar que todos os alunos que participaram do experimento estiveram em contato com o TuPy Online nas 2 semanas de aplicação.

A seguir apresentamos e analisamos os resultados obtidos com o experimento.

Numa primeira etapa, analisamos o questionário de caracterização dos estudantes pelas Figuras 15 a 23.

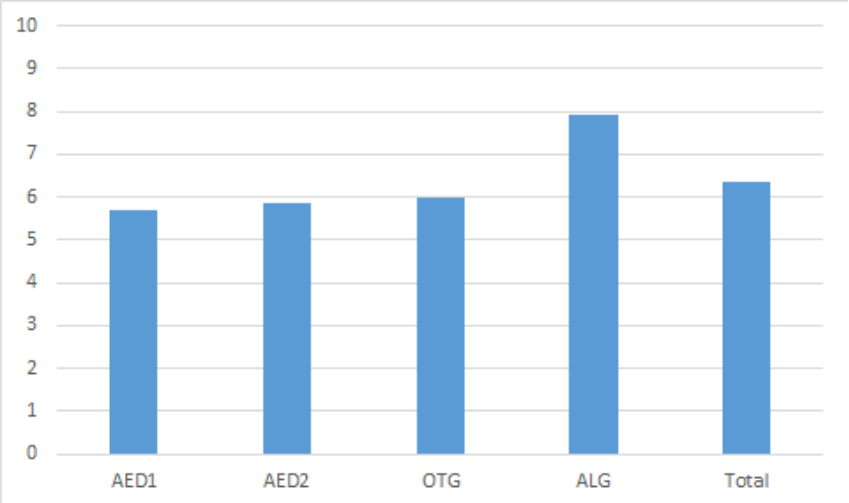
Figura 15 – Descrição das turmas por gênero.



Fonte: O próprio autor.

A Figura 15 mostra a distribuição da turma por gênero. Mais de 80% dos participantes foram do sexo masculino. O baixo número da presença feminina, impossibilita análises por gênero.

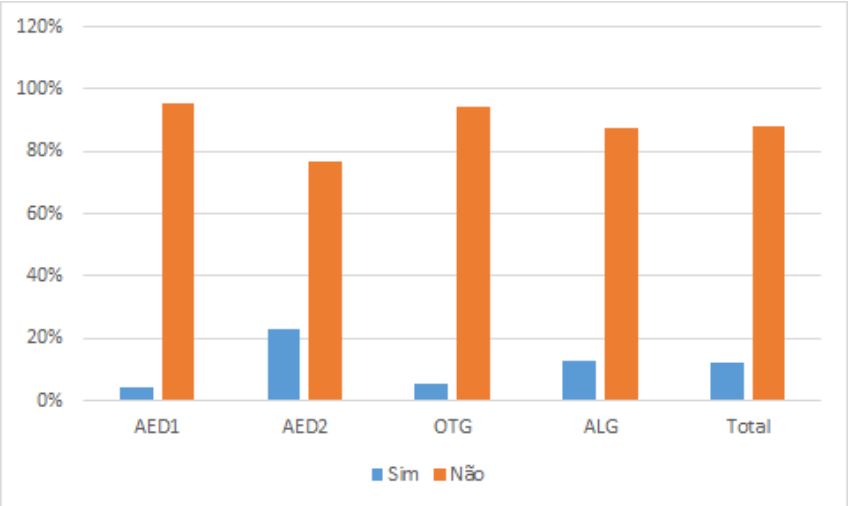
Figura 16 – Coeficiente de Rendimendo médio por Turma.



Fonte: O próprio autor.

A Figura 16 representa o coeficiente de rendimento (CR) médio entre os alunos de cada turma. O CR relatado pela turma do mestrado foi o obtido no final da graduação cursada pelo aluno. Verifica-se que, apesar da turma de AED1 possuir o menor CR, há uma similaridade com os CR das outras turmas de graduação. Já a turma do mestrado possui o maior CR, explicado pelo rígido processo seletivo.

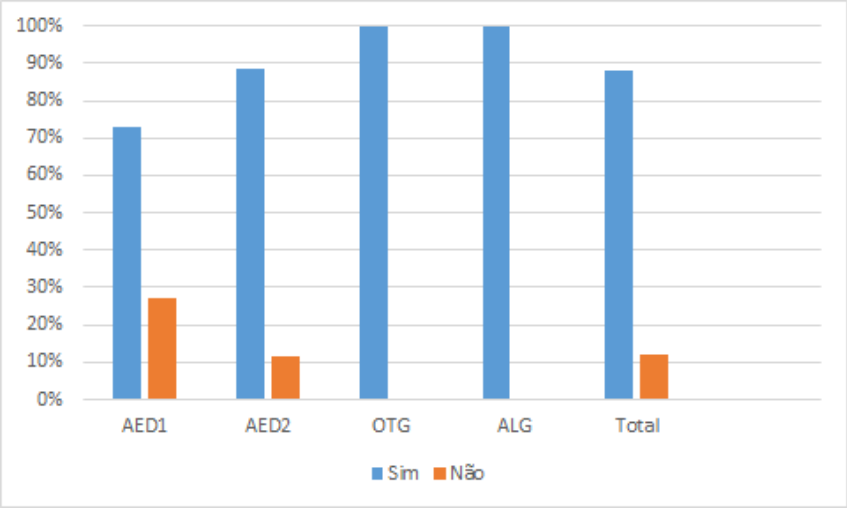
Figura 17 – Alunos que já cursaram a disciplina.



Fonte: O próprio autor.

A Figura 17 representa a distribuição dos alunos que já cursaram a disciplina. Mais de 70% em todas as turmas não haviam cursado a disciplina anteriormente. Este é um fato importante, pois o aprendizado com o TuPy Online poderia ter menor influência em alunos já expostos aos assuntos abordados.

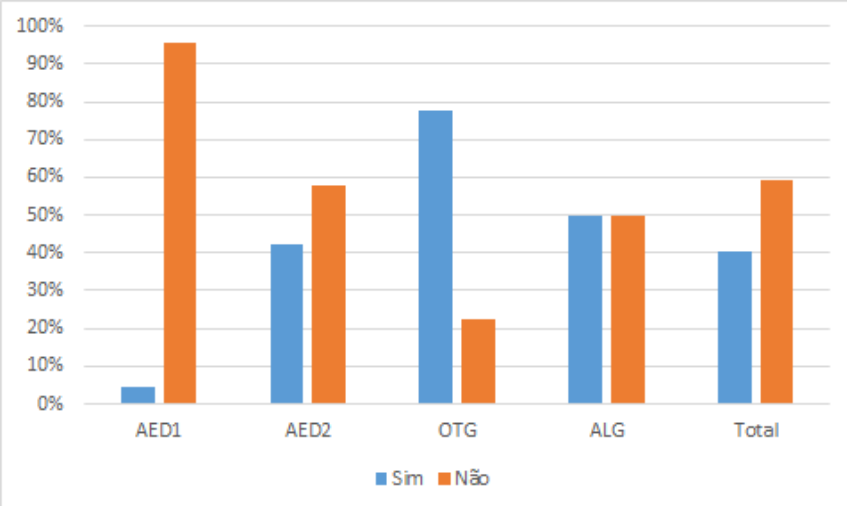
Figura 18 – Alunos com familiaridade em programação.



Fonte: O próprio autor.

A Figura 18 mostra a familiaridade dos alunos com programação. Nota-se que a frequência de alunos é maior que 70% e se apresentou de maneira uniforme.

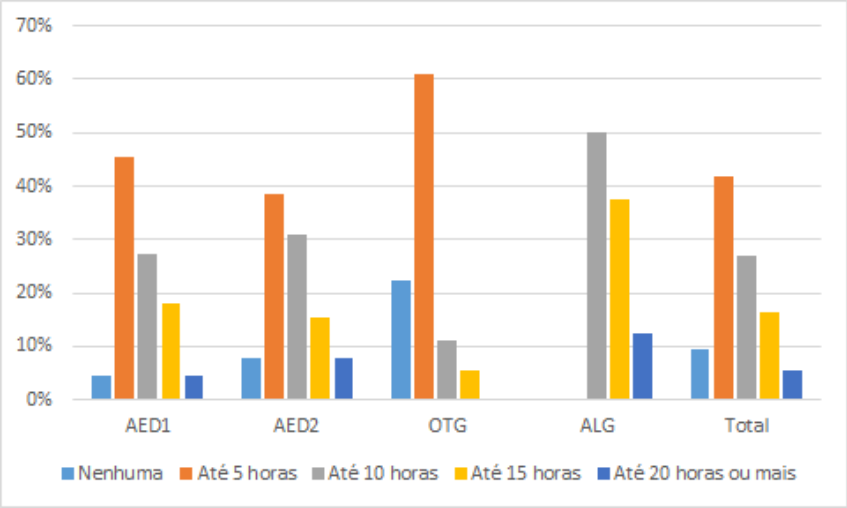
Figura 19 – Distribuição dos alunos que trabalham/estagiam por turma.



Fonte: O próprio autor.

A Figura 19 representa a frequência de alunos com trabalho ou estágio por turma. Verifica-se, dentre as turmas de graduação, que AED1 possui o menor percentual de alunos trabalhando e OTG possui o maior percentual. Isso é frequentemente normal na graduação, visto que AED1 é uma disciplina inicial no currículo acadêmico e as empresas buscam estagiários a partir da segunda metade do curso. Já AED2 e OTG vêm, respectivamente, depois. O mestrado, apesar da exigência na dedicação ao estudo apresentou 50% dos alunos com algum tipo de trabalho ou estágio.

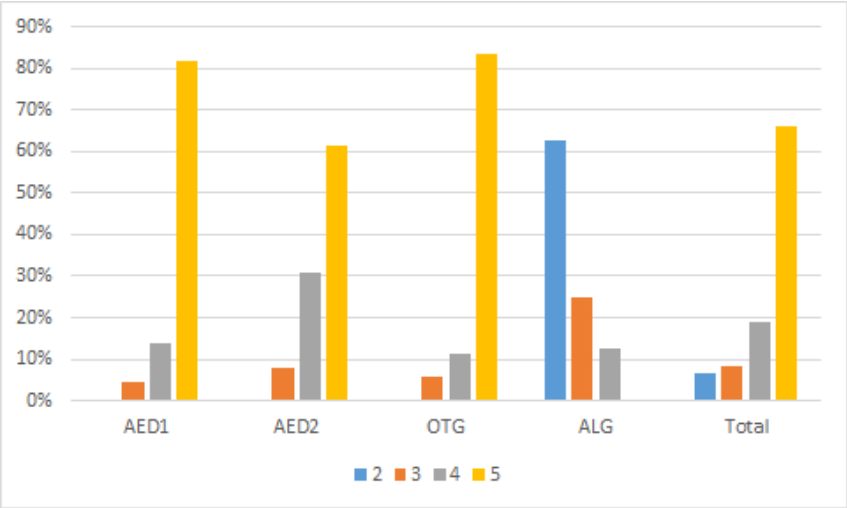
Figura 20 – Horas de estudo extra classe.



Fonte: O próprio autor.

A Figura 20, apresenta as horas de estudo extraclasse. Analisando os resultados, verifica-se que a distribuição de frequências de cada turma da graduação é bem similar, enquanto os alunos do mestrado relataram mais horas de estudo fora da sala de aula comparando-se com as outras turmas.

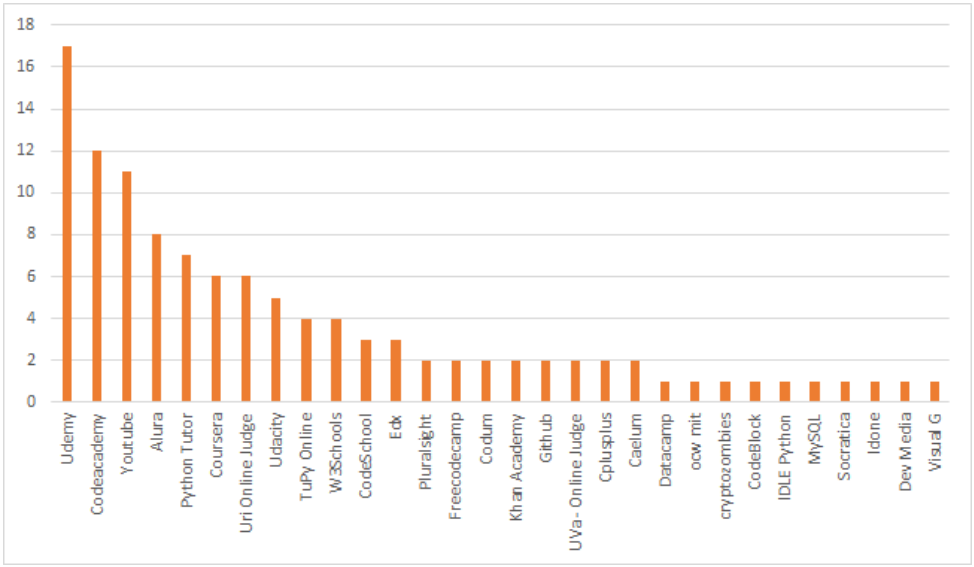
Figura 21 – Quantidade de disciplinas em curso no período da avaliação do TuPy Online.



Fonte: O próprio autor.

A Figura 21 representa a quantidade de disciplinas que o aluno participante do experimento estava cursando naquele período. As 3 turmas de graduação possuem comportamento semelhante. No entanto, não podemos comparar as turmas de graduação com a turma do mestrado, tendo em vista que a exigência de disciplinas na pós-graduação é elevada.

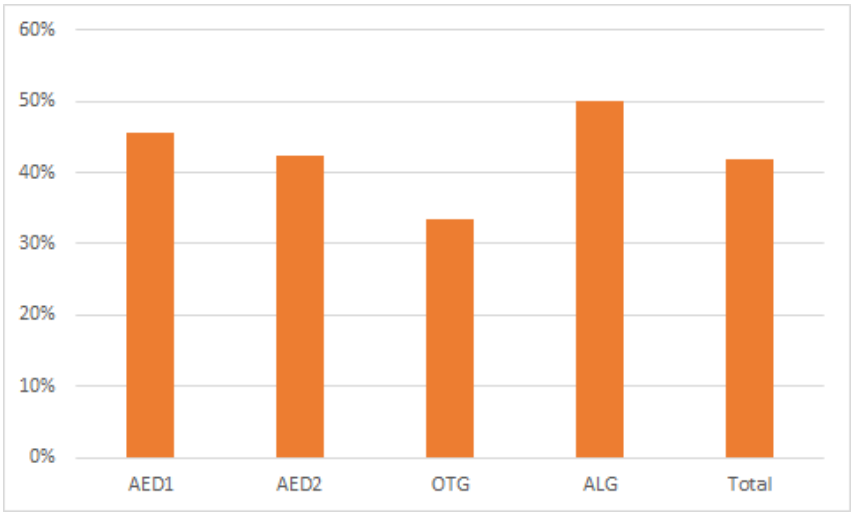
Figura 22 – Ferramentas de ensino de programação conhecidas pelos alunos.



Fonte: O próprio autor.

A Figura 22 representa todas as ferramentas de ensino de programação conhecidas pelos alunos. Vale ressaltar que foi permitido responder mais de uma ferramenta, além da liberdade da resposta. A *Udemy*, ferramenta de aprendizado online, foi a mais citada entre os alunos. A *Udemy* permite aos usuários criarem um curso, promovê-lo e ganharem dinheiro com as taxas de matrícula dos alunos. Além disso, *Codeacademy*, *Youtube*, *Alura*, *Online Python Tutor* formam o grupo das 5 ferramentas mais citadas. O Online Python Tutor, inclusive é a estrutura base para criação do TuPy Online. Dentre todos os alunos participantes, 9 citaram o TuPy Online como ferramenta de aprendizagem de programação, ou seja, já possuíam conhecimento prévio do programa a ser avaliado neste experimento.

Figura 23 – Alunos sem conhecimento de quaisquer ferramentas de ensino de programação.



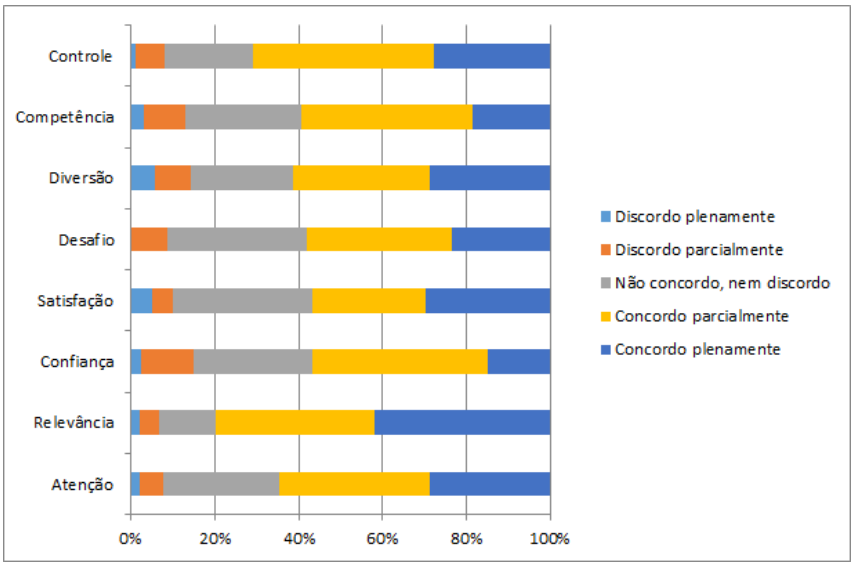
Fonte: O próprio autor.

Já a Figura 23 demonstra o percentual de alunos, por turma, que desconheciam quaisquer ferramentas de ensino. O resultado da turma do mestrado surpreende pelo percentual

alto de alunos sem conhecimento de ferramentas. Entretanto, isto se explica pelo fato de que muitos alunos não são oriundos dos cursos de computação na graduação.

A seguir, nas Figuras 24 e 25, e nas Tabelas 8 a 10, apresenta-se o resultado dos dados obtidos com o questionário de percepção.

Figura 24 – Avaliação do questionário de percepção.



Fonte: O próprio autor.

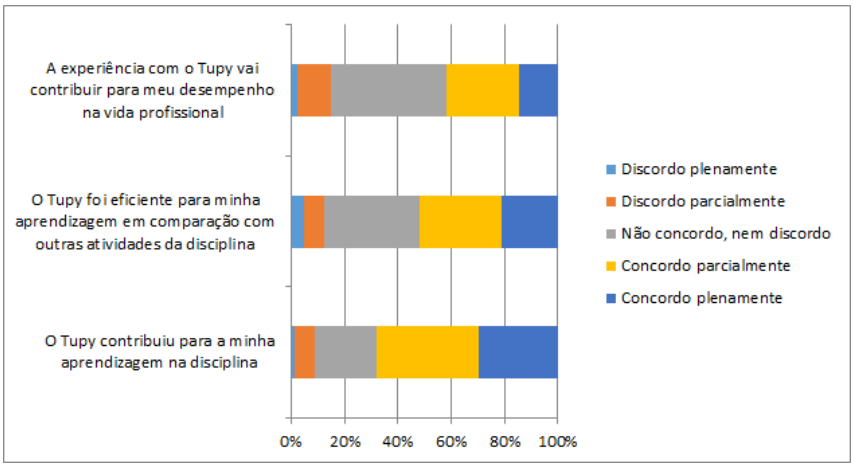
A Figura 24 refere-se às dimensões Motivação e Experiência do nível 1 de Kirkpatrick. Todos os critérios obtiveram índices de concordância (Concordo plenamente ou Concordo parcialmente) entre 55% a 80%, e de discordância (Discordo plenamente ou Discordo parcialmente) abaixo dos 20%, o que indica uma percepção altamente positiva em relação à ferramenta em análise.

Na dimensão Motivação, o item Relevância foi o que obteve maior avaliação. Em outras palavras, dentre os 4 componentes (Satisfação, Confiança, Relevância e Atenção), os alunos estão atribuindo grande importância aos conteúdos oferecidos pelo TuPy Online.

Sobre a Experiência do usuário, foram avaliados 4 itens (Desafio, Diversão, Competência e Controle). A dimensão Controle foi a melhor avaliada, ou seja, dada a experiência do aluno com o TuPy Online, a facilidade para usar os controles e interfaces do *software*, bem como as boas respostas para realizar ações obtiveram destaque perante aos usuários.

Quando questionados sobre a influência do TuPy Online no fator Aprendizagem (Figura 25), verificou-se que os alunos acreditam que o *software* tem grande importância e influência para o aprendizado na disciplina. No entanto, apesar da boa avaliação, não o consideram como uma ferramenta que contribua para o desempenho profissional. Tais percepções coincidem com os objetivos do TuPy Online, em ser uma ferramenta educacional no ensino de programação.

Figura 25 – Aprendizagem de curto e longo prazo.



Fonte: O próprio autor.

Tabela 8 – Resultados da avaliação pela Taxonomia de Bloom.

AED1	Lembrança		Compreensão		Aplicabilidade	
	Antes	Depois	Antes	Depois	Antes	Depois
Listas Lineares	2,61	3,00	2,55	3,38	2,11	2,61
Pilhas	3,00	3,44	2,50	3,22	2,27	2,72
Filas	3,06	3,44	2,77	3,38	2,50	2,83
OTG	**Lembrança**		**Compreensão**		**Aplicabilidade**	
	Antes	Depois	Antes	Depois	Antes	Depois
Busca em Grafos	2,66	3,12	2,66	3,20	2,37	3,12
Ordenação Topológica/COM	2,66	3,04	2,58	3,12	2,37	3,00
Componentes Fortemente Conexos	2,37	3,04	2,37	3,08	2,33	2,95
AED2	**Lembrança**		**Compreensão**		**Aplicabilidade**	
	Antes	Depois	Antes	Depois	Antes	Depois
Divisão e Conquista	3,00	3,55	2,88	3,77	2,66	3,33
Recursão e Memorização	3,33	3,77	3,55	3,77	3,11	3,44
Programação Dinâmica	3,44	3,77	3,55	3,77	3,11	3,44
ALG	**Lembrança**		**Compreensão**		**Aplicabilidade**	
	Antes	Depois	Antes	Depois	Antes	Depois
Listas Lineares	2,36	2,86	2,56	3,13	2,10	2,66
Pilhas	2,53	3,06	2,60	3,26	2,33	2,86
Filas	2,20	2,73	2,33	2,83	2,20	2,63

Fonte: O próprio autor.

A Tabela 8 apresenta os resultados da auto-avaliação dos alunos. Verificou-se que, para todos os assuntos abordados nas disciplinas, os alunos se auto-avaliaram melhores ao lembrar, compreender e aplicar a metodologia estudada após a utilização do TuPy Online. Tanto em AED1 quanto em ALG (mesmos assuntos apresentados nas disciplinas) os alunos relataram que o TuPy Online influenciou mais na compreensão do funcionamento de estrutura de listas lineares e pilhas. O pior desempenho encontrou-se na aplicabilidade e lembrança de filas. Em OTG, a aplicabilidade de busca em grafos e a compreensão de componentes fortemente conexos obtiveram as melhores auto-avaliações dos alunos. Por fim, em AED2, a compreensão em divisão e conquista obteve alto resultado. Os alunos também relataram que, em relação à compreensão em recursão, memorização e programação dinâmica, não perceberam melhorias após a utilização do TuPy Online.

Tabela 9 – Avaliação dos alunos em cada disciplina - Pré e Pós-teste.

Algoritmos e Estruturas de Dados I	Grupo 1		Grupo 2	
Alunos	Prova A	Prova B	Prova B	Prova A
1	8,3	8,1	1,9	5,3
2	4,9	0,2	1,2	1,0
3	3,2	6,5	7,7	7,3
4	6,1	7,0	5,8	7,1
5	6,8	7,0	5,7	3,2
6	5,9	7,4	2,7	1,2
7	1,7	0,1	4,1	6,7
8	4,0	4,4	0,1	0,1
9	7,5	7,2	9,4	9,0
10	7,3	4,8	7,0	6,3
11	9,0	9,1	2,5	4,3
12	0,5	0,9		

Otimização em Grafos	Grupo 1		Grupo 2	
Alunos	Prova A	Prova B	Prova B	Prova A
1	5,0	5,5	6,5	7,0
2	6,5	2,0	7,5	9,5
3	8,3	8,5	4,5	6,0
4	6,5	9,8	2,5	4,0
5	0,5	5,0	4,0	6,0
6	5,0	4,5	4,0	6,8
7	10,0	6,9	6,5	10,0
8	7,5	2,5	4,0	5,5
9	3,5	4,5	8,5	7,5
10	3,5	2,5	3,3	3,4
11	7,0	9,8		
12	2,0	8,5		

Algoritmos e Estruturas de Dados II	Grupo 1		Grupo 2	
Alunos	Prova A	Prova B	Prova B	Prova A
1	4,5	4,0	3,9	8,0
2	5,0	4,5	3,0	4,8
3	5,0	7,0	4,5	6,8
4	7,0	5,5	2,0	2,3
5	1,5	0,5	2,0	3,2
6	4,3	3,3	2,3	6,0
7	4,3	2,8	3,5	7,5
8	5,7	6,9	3,0	4,5
9	2,0	2,0	3,3	3,0
10	4,5	2,3	2,3	6,0
11	6,5	3,5	2,3	4,7
12	6,2	6,5	4,3	5,5
13	3,5	3,8	3,0	2,3
14	1,8	3,5		
15	9,1	9,3		

Algoritmos - Mestrado	Grupo 1		Grupo 2	
Alunos	Prova A	Prova B	Prova B	Prova A
1	6,7	7,6	9,9	9,8
2	7,0	8,3	8,1	8,4
3	7,4	8,1	5,2	5,7
4	5,4	9,9	6,2	4,5
5	6,4	7,5	7,7	8,9
6			9,5	9,7
7			7,6	9,6

Fonte: O próprio autor.

A Tabela 9 apresenta os resultados do pré-teste e pós-teste e a a Tabela 10 representa as médias e medianas destes testes. Na avaliação de aprendizagem pelas notas obtidas realizou-se inicialmente o teste de normalidade dos dados (Através do *software* Statistica 12) e, considerando que as amostras são dependentes, foi realizado o t pareado, em caso de normalidade da distribuição, e o teste de Wilcoxon pareado [43], em casa de não normalidade. Para os grupos de controle, foi utilizado o teste U de Mann-Whitney. Para testes com tamanhos de amostra maiores que 25 (valores em que os testes de Wilcoxon pareado e Mann-Whitney se aproximam da normal), foi utilizado o teste t pareado. A hipótese nula adotada é a não influência do TuPy Online. A hipótese alternativa é a melhora no desempenho das notas rem relação ao pré-teste, sendo assim utilizaremos um teste de hipótese unilateral à direita. Os resultados são apresentados a seguir.

Tabela 10 – Médias e Medianas da avaliação dos alunos em cada disciplina - Pré e Pós-teste.

Algoritmos e Estruturas de Dados I	Grupo 1		Grupo 2		Otimização em Grafos	Grupo 1		Grupo 2	
Alunos	Prova A	Prova B	Prova B	Prova A	Alunos	Prova A	Prova B	Prova B	Prova A
Média	5,43	5,23	4,37	4,68	Média	5,80	3,50	4,00	6,60
Mediana	6,00	6,75	4,10	5,30	Mediana	5,00	3,50	4,00	6,00
Algoritmos e Estruturas de Dados II	**Grupo 1**		**Grupo 2**		**Algoritmos - Mestrado**	**Grupo 1**		**Grupo 2**	
Alunos	Prova A	Prova B	Prova B	Prova A	Alunos	Prova A	Prova B	Prova B	Prova A
Média	4,73	4,36	3,03	4,97	Média	6,58	8,28	7,74	8,09
Mediana	4,50	3,80	3,00	4,80	Mediana	6,70	8,10	7,70	8,90

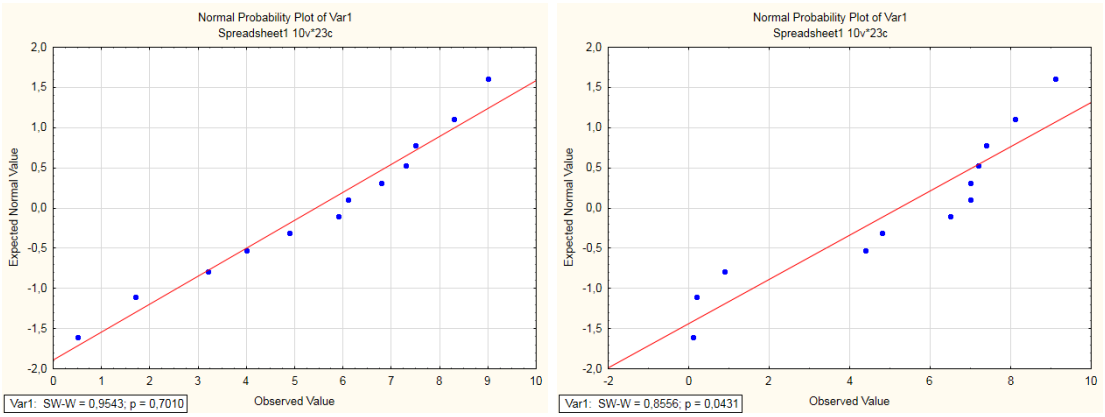
Fonte: O próprio autor.

Considerando todos os alunos de todas as disciplinas, utilizando o teste t pareado, em que 85 alunos participaram, a análise indicou a rejeição da hipótese nula para um nível de confiança de 95%, para o p-valor de 0,00506 (menor nível de significância para rejeitar H_0). Em outras palavras, houve uma melhora significativa no aprendizado dos alunos com a utilização do TuPy Online.

As turmas de AED1 e ALG eram ministradas pelo mesmo professor. Sendo assim, fez-se uma avaliação conjunta de todas as notas dos alunos destas turmas. Neste caso, utilizou-se novamente o teste t pareado. O p-valor encontrado foi de 0,1345. Ou seja, para um nível de confiança de 95%, avaliando as turmas deste mesmo professor, não encontrou-se evidências estatísticas que fizessem rejeitar a hipótese nula.

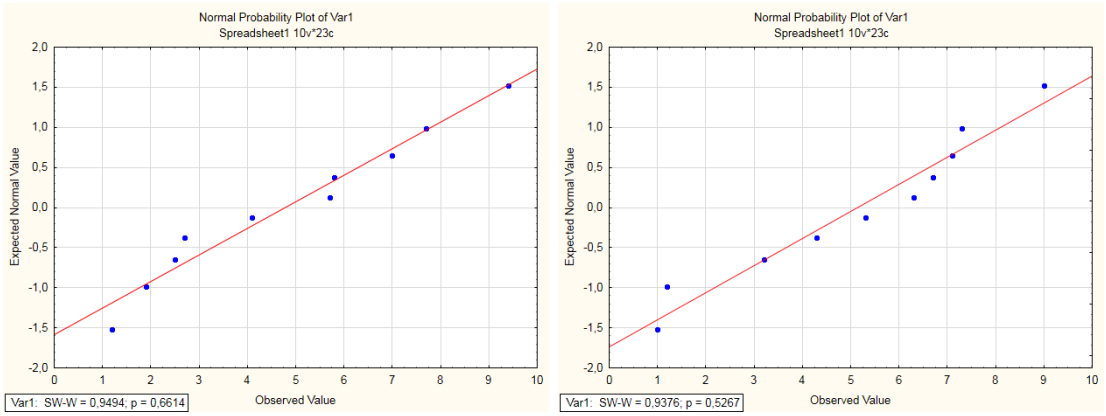
Da mesma forma, em AED2 e OTG, turmas de outro professor, foram avaliadas as notas conjuntas obtidas pelos alunos destas disciplinas. Neste caso, o p-valor encontrado foi de 0,0099. Ou seja, rejeita-se a hipótese nula para um nível de confiança de 95%, pois encontrou-se evidências para acreditar que o TuPy Online auxiliou no desempenho do aluno.

Figura 26 – Grupo 1 - AED1 - Pré-teste e Pós-teste.



Fonte: O próprio autor.

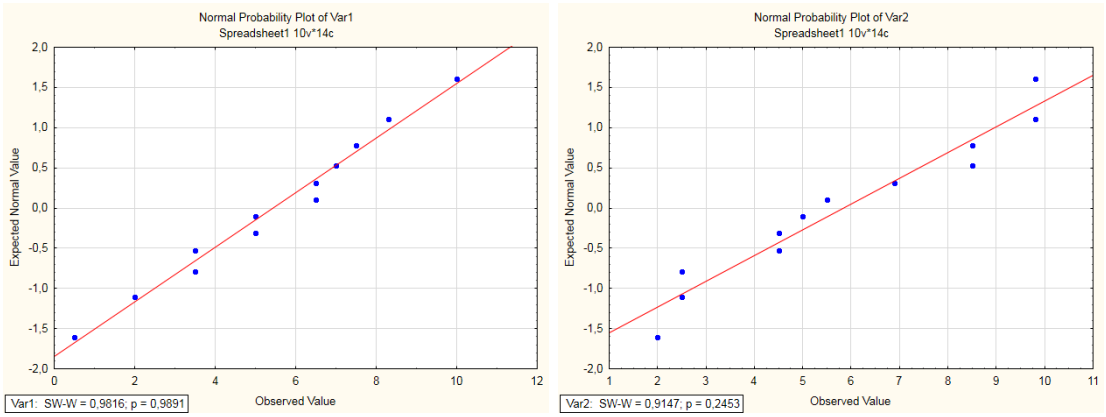
Figura 27 – Grupo 2 - AED1 - Pré-teste e Pós-teste.



Fonte: O próprio autor.

Na disciplina AED1, para o Grupo 1, analisando a Figura 26, verifica-se normalidade nos dados para as notas do pré-teste. Em relação ao pós teste, tanto a técnica gráfica quanto o teste de Shapiro-Wilk (p-valor $< 5\%$) traduzem uma interpretação de não normalidade. Sendo assim, para este Grupo, utilizou-se o teste de Wilcoxon pareado. Em relação ao Grupo 2, vide Figura 27, tanto o pré-teste quanto os dados do pós-teste indicaram normalidade possibilitando a utilização do teste t pareado. Sendo assim, a análise intragrupo resultou em um valor observado de 37,5 para o Grupo 1 frente ao valor crítico de Wilcoxon (17) não possibilitando a rejeição da hipótese nula, ou seja, sem evidências estatísticas para comprovar que o TuPy Online influenciou no desempenho dos alunos. Obteve-se resultado análogo para o Grupo 2, com p-valor igual a 0,28962.

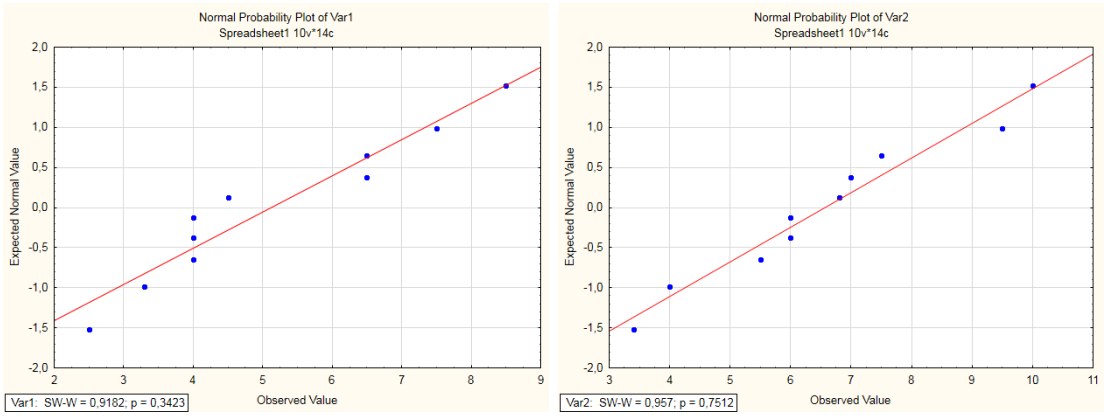
Figura 28 – Grupo 1 - OTG - Pré-teste e Pós-teste.



Fonte: O próprio autor.

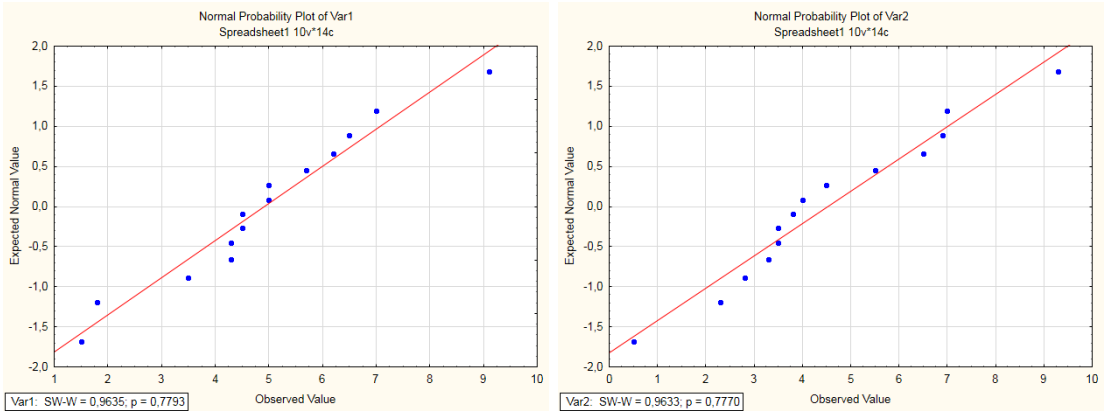
Na disciplina de OTG, analisando as Figuras 28 e 29, verificou-se normalidade dos dados, através da técnica gráfica e do teste de Shapiro-Wilk. Sendo assim, para análise de pré-teste e pós-teste dos 2 Grupos, utilizou-se o teste t pareado. O teste t pareado rejeitou a hipótese nula para o Grupo 2, com p-valor de 0,0034. No entanto, para o Grupo 1, não houve rejeição da hipótese nula, com p-valor observado de 0,3539. Para o Grupo 2, há indicação que o mesmo foi influenciado positivamente pelo TuPy Online.

Figura 29 – Grupo 2 - OTG - Pré-teste e Pós-teste.



Fonte: O próprio autor.

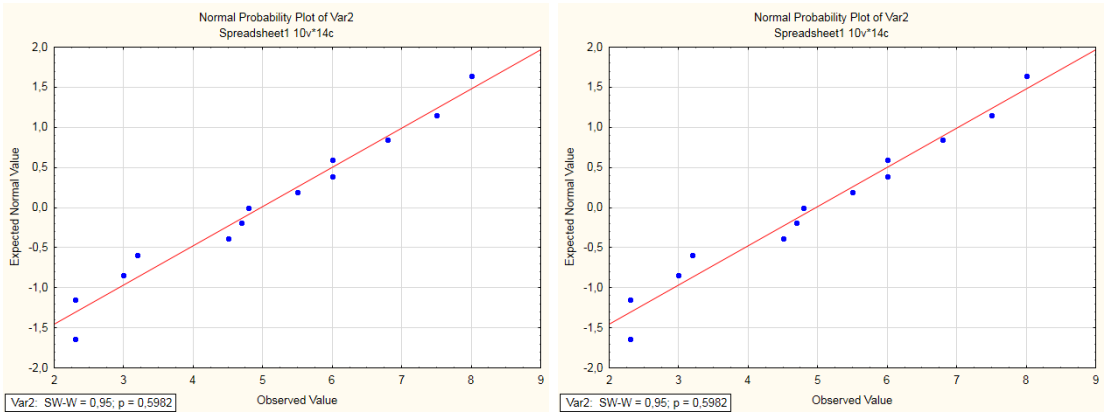
Figura 30 – Grupo 1 - AED2 - Pré-teste e Pós-teste.



Fonte: O próprio autor.

Na disciplina de AED2, também constatamos a normalidade dos dados nos 2 Grupos, Figuras 30 e 31. Possibilitando a utilização do teste t pareado. Inclusive o teste de Shapiro-Wilk, extremamente rígido, não rejeitou a hipótese de normalidade, vide os p-valores encontrados (canto esquerdo inferior de cada gráfico). Assim como em OTG, só o Grupo 2 representou melhoria significativa, ou seja, através do teste t pareado, encontrou-se evidências para rejeição da hipótese nula. No Grupo 1 encontrou-se p-valor de 0,1642. Para o Grupo 2 o p-valor foi de 0,0005. Como o p-valor observado para o Grupo 2 foi menor 5%, rejeita-se a hipótese nula sob o nível de 95% de confiança.

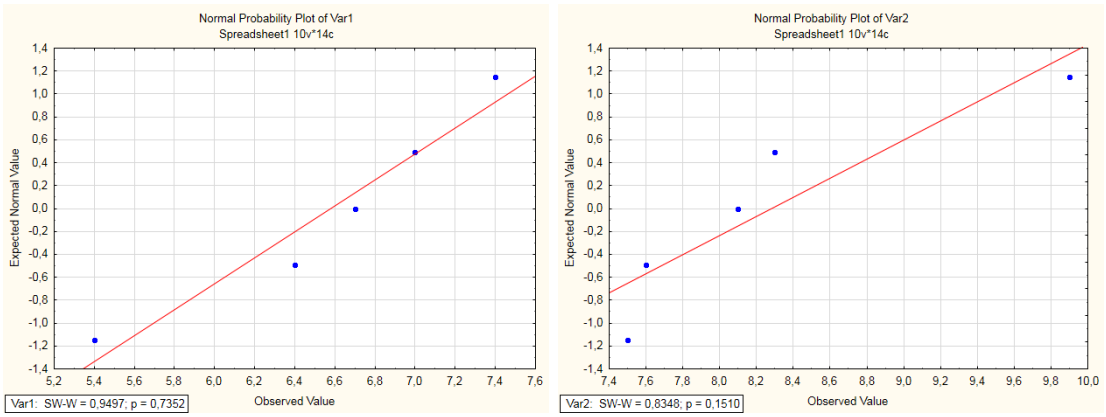
Figura 31 – Grupo 2 - AED2 - Pré-teste e Pós-teste.



Fonte: O próprio autor.

Na disciplina de ALG, avaliando a normalidade dos dados, Figuras 32 e 33, concluise que no Grupo 1 há normalidade dos dados, já no Grupo 2 não há normalidade nos dados do pós-teste, verificado pela técnica gráfica e confirmado pelo teste de Shapiro-Wilk. O TuPy demonstrou melhora apenas em um grupo, rejeitando a hipótese nula. No Grupo 1 obteve-se p-valor de 0,0370. No Grupo 2, o valor observado no teste de Wilcoxon foi de 7 com valor crítico de 4. Neste grupo, rejeita-se a hipótese nula, contudo, se desconsiderarmos os alunos que reportaram não terem estudado em casa com o TuPy Online. Novamente utilizou-se 95% de confiança.

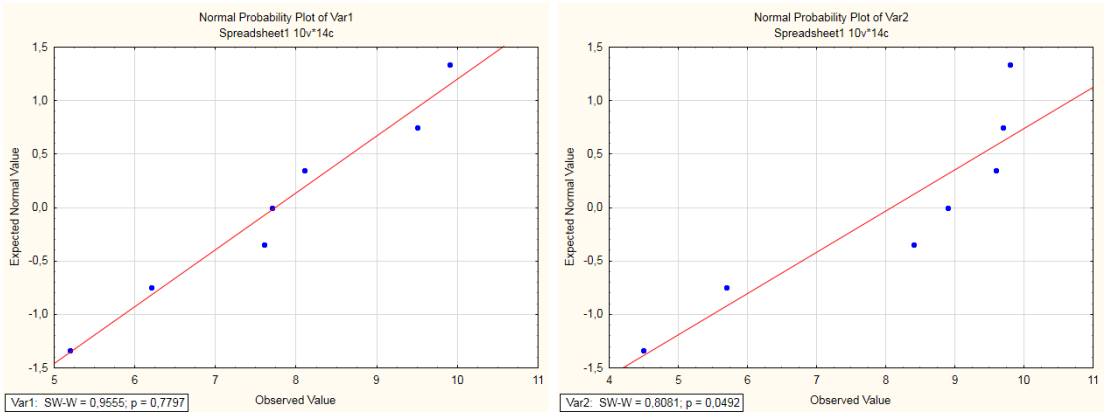
Figura 32 – Grupo 1 - ALG - Pré-teste e Pós-teste.



Fonte: O próprio autor.

Em relação aos grupos de controle, criados para estudar o efeito das provas no desempenho dos alunos, o resultado foi inconclusivo em mostrar que elas não tiveram influência. Em todos os casos, o Teste de Mann-Whitney não encontrou evidências para rejeitar a hipótese nula, isto indica que, não pode-se apontar melhorias de desempenho com o uso do TuPy Online na realização da mesma prova, ora aplicada para um Grupo antes do uso do TuPy Online, ora aplicada para o outro Grupo após a utilização do TuPy Online. Sendo assim, há a possibilidade de que a melhoria observada seja parcialmente devida à provas com níveis de dificuldades distintas. Os testes de U de Mann-Whitney foram realizados sob nível de confiança de 95%.

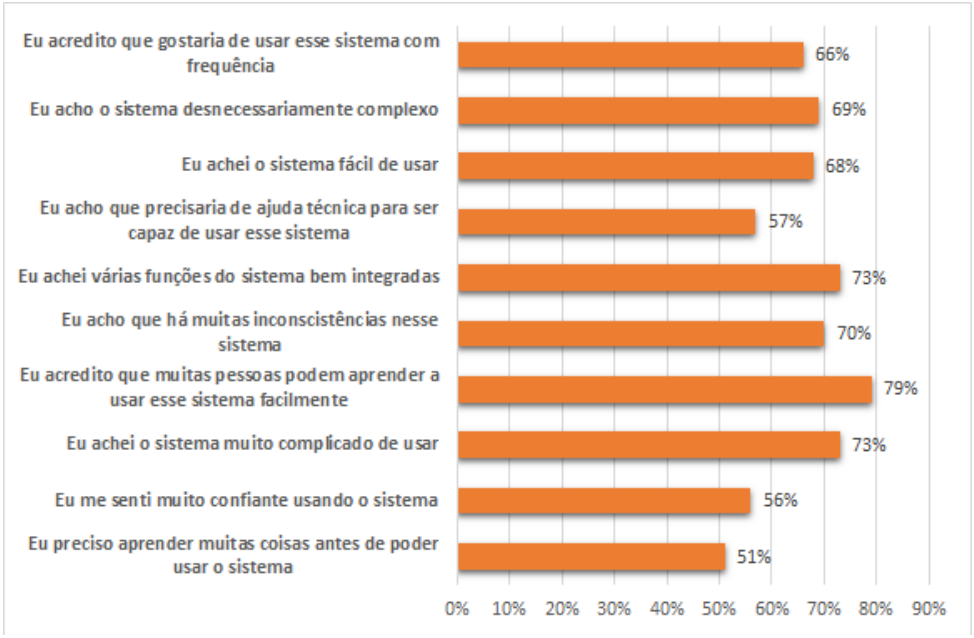
Figura 33 – Grupo 2 - ALG - Pré-teste e Pós-teste.



Fonte: O próprio autor.

Um primeiro resultado da análise sobre usabilidade, considerando as respostas de maneira global, foi o percentil de respostas positivas encontrado, de 66%. Comparando com a análise feita por [42], o resultado é bem próximo à média de 68%, relativa a cerca 500 análises. Essa média está homogeneamente distribuída entre os ítens analisados, sendo o de pior avaliação a satisfação, com 60% (questões 1, 4, 9) do total de pontos possíveis e o melhor avaliado a eficiência, com 72% (questões 5, 6, 8). Esses resultados sugerem que os usuários consideram as funcionalidades do *software* bem integradas, de execução rápida, mas ainda necessitam de ajuda técnica para o uso da ferramenta. Foram dadas inúmeras sugestões de melhorias, que serão consideradas em futuras versões. Vide Figura 34. Além disso a facilidade de aprendizagem (questões 3, 4, 7, 10), facilidade do usuário em aprender a usar o sistema, obteve média de 64%. A facilidade de memorização (questão 2), facilidade em utilizar o sistema após um período de não uso, obteve 69%. E por fim, minimização dos erros (questão 6), avaliando as inconsistências no sistema, obteve 70%.

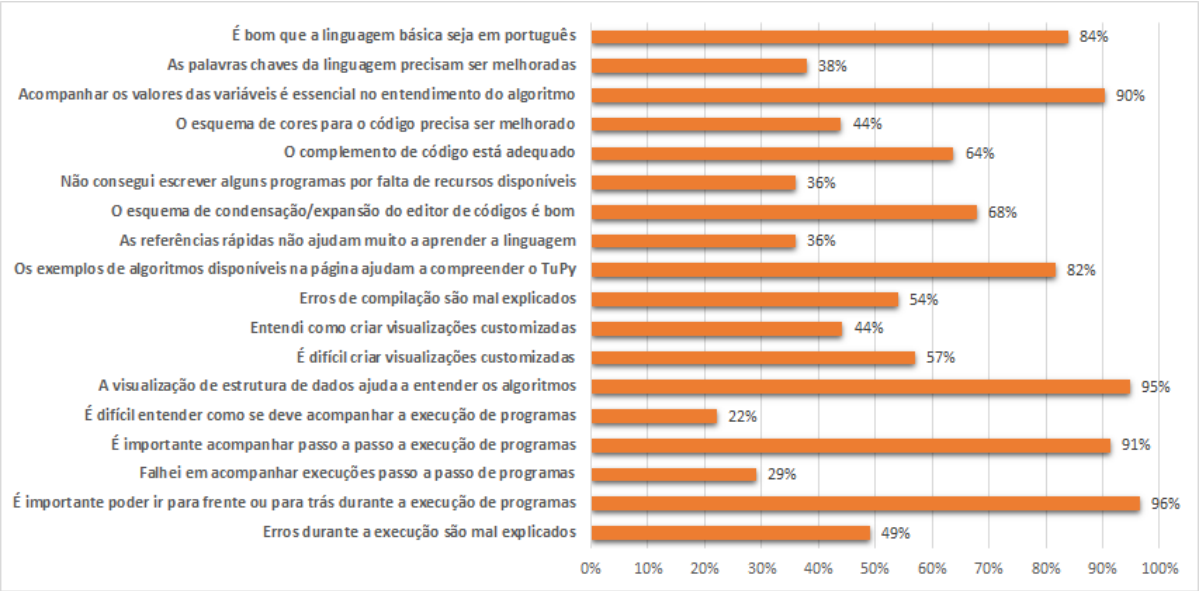
Figura 34 – Avaliação de Usabilidade SUS - questão por questão.



Fonte: O próprio autor.

Em relação ao questionário de *feedback* dos alunos, os resultados serão dispostos da seguinte forma. Primeiramente os resultados obtidos através da escala Likert das questões de múltipla escolha em que o participante marcaria um valor de 0 a 4, indicando desde "Discorda fortemente" até "Concorda fortemente". Em seguida, apresenta-se o gráfico com os algoritmos mais simulados pelos alunos. E, por fim, as opiniões livres dos alunos referentes aos aspectos da linguagem, interface e processo introdutório do TuPy Online.

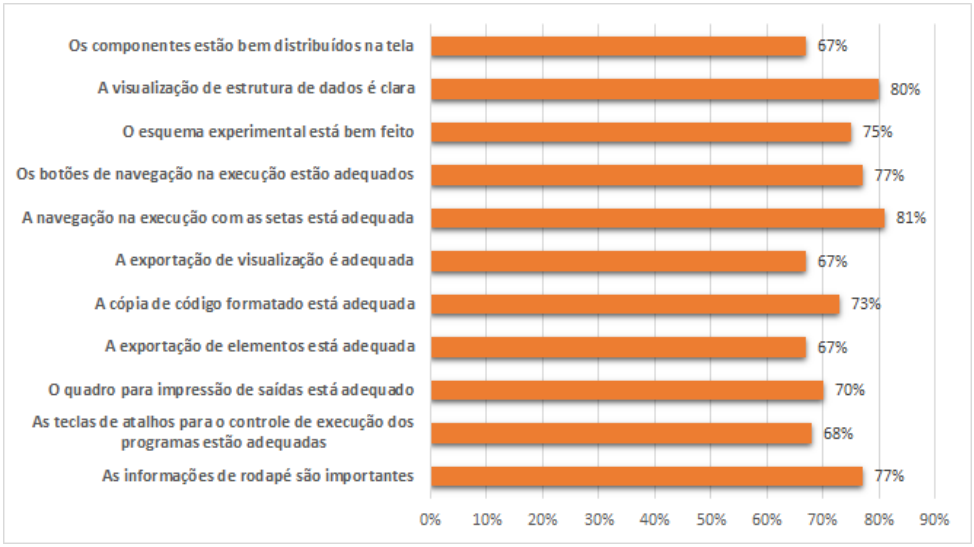
Figura 35 – *Feedback* - Aspectos de Linguagem.



Fonte: O próprio autor.

Na Figura 35, apresentam-se os resultados das respostas dos alunos aos aspectos da linguagem do TuPy Online. As perguntas ímpares, são perguntas positivas, ou seja, as que possuírem maior percentual são qualidades e as de menor percentual as que necessitam de melhorias. Dentre as perguntas positivas, a possibilidade de poder avançar ou retroceder o programa durante a execução obteve 96% da máxima pontuação possível. O ponto negativo, neste caso, é o entendimento em criar visualizações personalizadas dentro do algoritmo, este aspecto obteve apenas 44%. Em relação as perguntas pares, negativas, analisa-se de forma contrária. As perguntas que receberam percentual alto são potenciais problemas encontrados pelos alunos. As questões que receberam percentual baixo, são qualidades encontradas no *software*. Neste caso, somente 2 perguntas apresentaram um percentual maior que 50%. A dificuldade em criar visualizações (57%) e erros de compilação mal explicados (54%), representaram a maior preocupação dos alunos em relação aos aspectos da linguagem no TuPy Online. Os pontos fortes, foram a facilidade em acompanhar a execução de programas (22%) e o acompanhamento na execução passo a passo (29%).

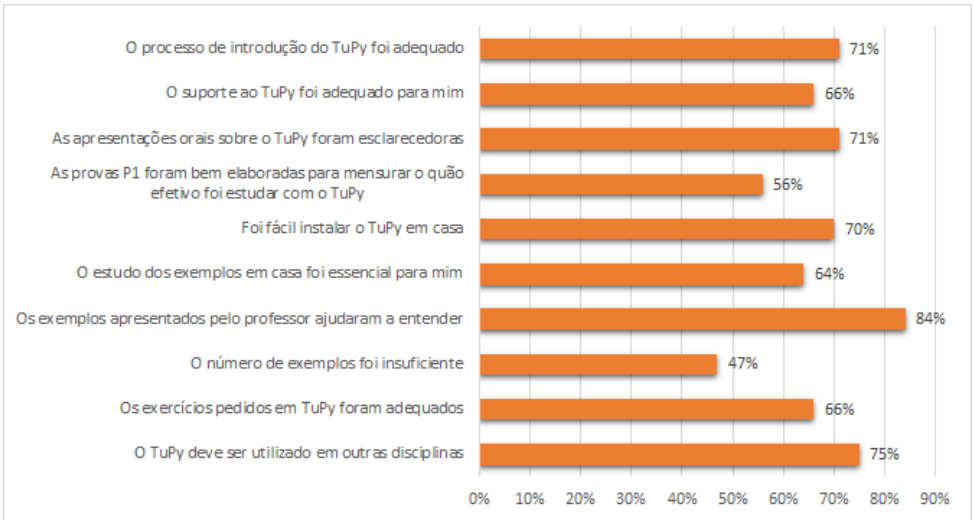
Figura 36 – *Feedback* - Interface.



Fonte: O próprio autor.

Na Figura 36 não há perguntas negativas, sendo assim a análise é direta com os percentuais obtidos. Em relação à interface e suas funcionalidades, os alunos indicaram que a navegação na execução com as setas (81%) e a clareza na visualização de estrutura de dados (80%) são os pontos fortes. A exportação de elementos (67%), a exportação de visualização (67%) e a distribuição dos componentes na tela (67%) apesar de bem avaliados, foram os pontos que receberam menor avaliação dentre estas 11 questões e poderiam ser possíveis candidatos a melhorias.

Figura 37 – *Feedback* - Apresentação Introdutória.

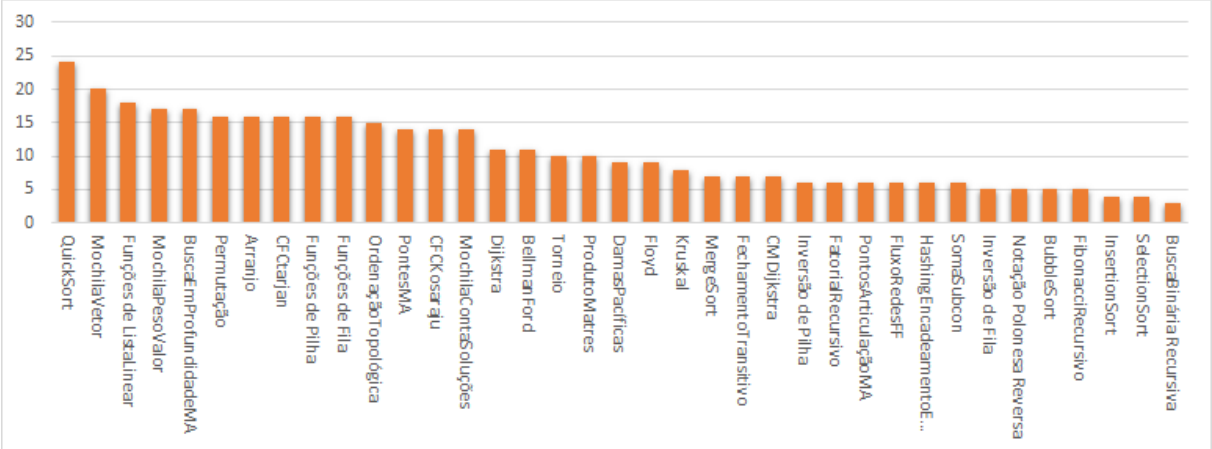


Fonte: O próprio autor.

Na Figura 37 os alunos avaliaram o processo introdutório do TuPy Online realizado pelos professores de cada disciplina. Há uma pergunta negativa, ou seja, ela será analisada de maneira inversa como anteriormente. Os alunos mostraram-se satisfeitos em relação ao processo introdutório realizado pelos professores para o experimento com o TuPy Online. Os exemplos apresentados, na opinião da maioria dos alunos, ajudaram a entender melhor

software. O item com menor avaliação dos alunos, foi no aspecto da elaboração das provas (56%) para efeito de pré-teste e pós-teste na avaliação de aprendizagem.

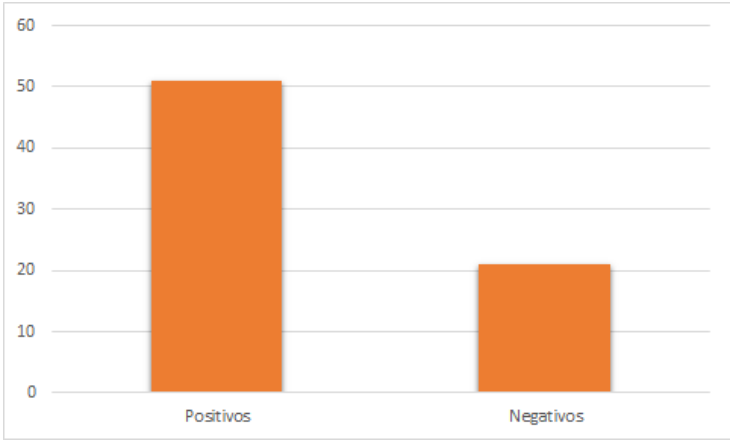
Figura 38 – Número de alunos que simularam determinado algoritmo.



Fonte: O próprio autor.

Na Figura 38 apresenta-se os algoritmos que foram sugeridos em sala para simulação e a quantidade de alunos que os simulou. Vale ressaltar que todos os algoritmos foram simulados ao menos por um aluno. Os algoritmos mais simulados foram: *Quicksort*, MochilaVetor, Funções de Listas Lineares, MochilaPesoValor e Busca em Profundidade.

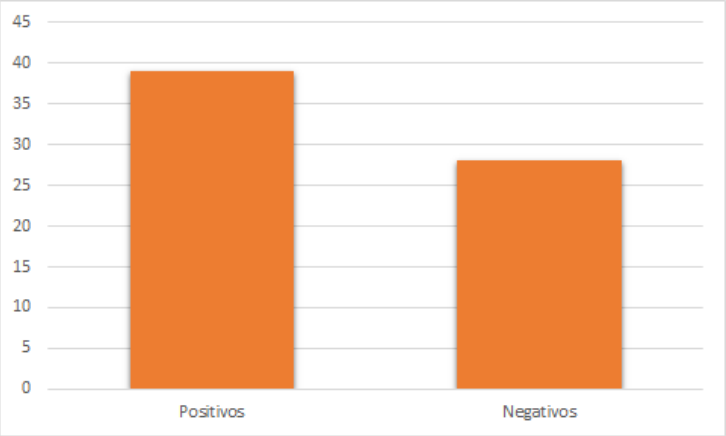
Figura 39 – Comentários livres sobre os aspectos da linguagem.



Fonte: O próprio autor.

Na Figura 39, mostra-se que mais alunos fizeram comentários positivos que negativos referente ao aspectos de linguagem. Em geral, os comentários positivos foram acerca de um sistema intuitivo, fácil, didático e com linguagem em português, com linguagem simples, direta e possibilitando uma visualização simplificada do algoritmo. Outro ponto bastante comentado positivamente foi a possibilidade de executar o programa passo a passo além de implementações sem grande dificuldades. Em relação aos pontos negativos, a maioria criticou na pouca explicação dentro do programa de como realizar visualizações customizadas. Além disso, foi criticada a organização de visualizações complexas na tela e erros reportados de maneira pouco clara.

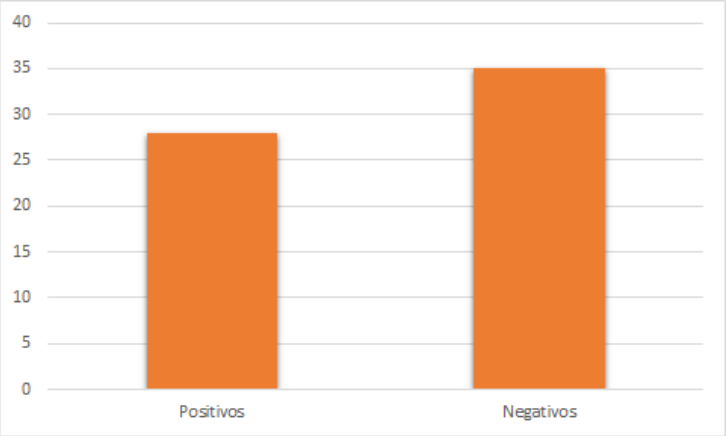
Figura 40 – Comentários livres sobre a interface.



Fonte: O próprio autor.

Na Figura 40, novamente, há mais comentários positivos que negativos acerca da interface. No entanto, esta diferença é mais discreta em relação aos aspectos da linguagem. Em geral, os comentários positivos referem-se a interface simples, limpa, bem integrada, leve, objetiva e funcional. Já os comentários negativos referem-se ao enquadramento na tela das visualizações, botões com maior destaque na tela, problemas na edição do código após o início do algoritmo e problemas com letras maiúsculas no algoritmo.

Figura 41 – Comentários livres sobre o processo introdutório.



Fonte: O próprio autor.

Por fim, os comentários em relação ao processo introdutório estão representados na Figura 41. Neste caso, há mais comentários negativos que positivos. A maioria dos comentários positivos são referentes as aulas e explicações dos professores com clareza e objetividade em transmitir o máximo de conhecimento acerca do TuPy Online. Já os pontos negativos, em geral, referem-se ao pouco tempo do experimento, a falta de exemplos mais básicos para o aprendizado da linguagem TuPy e possibilidade de um maior suporte ao programa.

CONCLUSÃO

Trabalhar com novas tecnologias auxiliando o ganho de aprendizagem tem-se tornado frequente nos últimos tempos, apesar de não trivial. Além disso, é primordial não só trabalhar, mas avaliar se tais metodologias influenciam positivamente no ensino do aluno. O TuPy Online, ferramenta ancorada nas bases do Online Python Tutor, na personalização visual do *Graphviz* e criado para auxiliar a aprendizagem em cursos de graduação, necessitava de uma avaliação de aprendizagem a fim de ratificar sua importância no ensino.

Com isso, este trabalho garimpou metodologias de análises da eficácia de ferramentas no ganho de aprendizagem e de avaliação de usabilidade. Ponderou suas limitações, peculiaridades e características de avaliação. Optou-se, então, para avaliar a aprendizagem, por utilizar a metodologia proposta por Savi et al. que possui base metodológica nos 4 níveis de treinamento de Kirkpatrick. Além da contribuição proposta por Ferreira em um estudo posterior. Para realização do teste de usabilidade, optou-se pelo questionário SUS, além de uma complementação de um questionário de *feedback* para um retorno específico de melhorias para o *software*.

O estudo realizado na UERJ, contou com a presença maciça dos alunos inscritos nas disciplinas selecionadas, participando das etapas do experimento, principalmente das semanas em que o TuPy Online foi apresentado.

A avaliação do TuPy Online através do questionário de percepção demonstrou um alto índice de concordância em relação aos aspectos de relevância em seu conteúdo, bem como a facilidade e praticidade de sua interface e seus controles internos. Inclusive, tais índices, obtiveram resultados similares aos encontrados por Savi [3] e Ferreira [25].

Ainda sobre o questionário de percepção, no pilar motivacional descrito pelo modelo de Keller [14], o quesito Confiança obteve o resultado menos satisfatório. Ou seja, os alunos indicaram ausência de sensações de progresso durante a utilização do TuPy Online. Um possível direcionamento do TuPy Online para uma determinada disciplina incluíndo uma sequência de algoritmos com níveis de dificuldade crescentes para exercícios poderia auxiliar em uma melhoria na avaliação deste quesito.

O estudo é conclusivo para avaliação do TuPy Online. Enquanto que em AED1 os testes estatísticos não indicam melhora significativa para o desempenho nas duas provas, nas disciplinas de OTG, AED2 e ALG, principalmente para este último, houve melhora significativa. Importante indicar que a turma com menor CR foi a única que não obteve melhora no desempenho no pré-teste e pós-teste.

Uma possibilidade para explicar o melhor desempenho dos alunos mestrandos, é que, notoriamente, esses estudantes dedicaram um número maior de horas de estudo extra-classe, podendo assim, ter contribuído para uma avaliação mais realista da ferramenta.

Outra observação importante é que o TuPy Online obteve significância de aprendizagem nas turmas em que o conhecimento de programação era baixo (o público do mestrado é formado por graduados de diversas áreas que não de computação) ou em disciplinas mais avançadas, onde as visualizações de estruturas de dados em diversos níveis de abstração realmente podem ser mais decisivas. O estudo sugere que as disciplinas intermediárias, que não requerem visualização avançada e que já pressupõem um nível básico de progra-

mação, não são significativamente beneficiadas pelo TuPy Online.

De qualquer forma, apesar de alguns grupos não demonstrarem um melhor desempenho, a partir do contato com a ferramenta, a auto-avaliação e o retorno satisfatório por parte dos alunos de todas as turmas reiteram a importância do TuPy Online para aproximar a disciplina ao aluno.

Em relação à usabilidade, para um primeiro teste, o resultado foi satisfatório. O TuPy Online obteve um índice próximo da média das avaliações de outros estudos com o mesmo questionário de usabilidade SUS. Dentre as características de usabilidade, as funcionalidades bem integradas e de execução rápida obtiveram destaque.

Em relação ao questionário e aos comentários de *feedback* dos alunos o resultado foi positivo. Entre todos os pontos, o que mais agradou foi a possibilidade de avançar ou retroceder passo a passo. Outros pontos positivos foram a utilização da língua portuguesa e o modelo de visualizações de algoritmos. No entanto, as visualizações customizadas foram alvo de críticas no que diz respeito ao auxílio de como proceder.

Pontos de interseções importantes na avaliação de aprendizagem através da utilização das provas e a avaliação dos alunos foram importantes. O tempo de utilização do TuPy Online (2 semanas), considerado como possível hipótese para o desempenho do nível 2 de Kirkpatrick ter produzido um resultado não tão bom quanto o nível 1, foi confirmado também pelos alunos. Em geral, as respostas indicaram que as 2 semanas foi um curto período para ambientação do *software*.

A partir desta experiência na avaliação do TuPy Online para algumas turmas de Computação pode-se indicar alguns outros estudos de utilidade. Considerar um período maior para a introdução do *software*, poderiam validar melhor a ferramenta, visando esclarecer alguns resultados inconclusivos. Além disso, poderia ser interessante avaliar a introdução da mesma em outros cursos universitários além da Ciência da Computação. Fora isso, há também, como proposta para trabalhos futuros, produzir uma nova versão do *software* considerando o *feedback* positivo e construtivo através da opinião dos alunos.

Os resultados parciais deste trabalho foram apresentados no 27º Workshop em Educação em Informática (WEI) no Congresso da Sociedade Brasileira de Computação (CSBC) de 2019 [48] e no 5º Workshop de Ensino em Pensamento Computacional, Algoritmos e Programação (WAlgProg) no VIII Congresso Brasileiro de Informática na Educação (CBIE) de 2019 [49].

REFERÊNCIAS

[1] ARRUDA, J. *Modelagem do Processo de Aprendizagem na Educação Superior*. Rio de Janeiro: edUERJ, 2007.

[2] FU, F.; SU, R.; YU, S. Egameflow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education*, Elsevier, v. 52, n. 1, p. 101–112, 2009.

[3] SAVI, R.; WANGENHEIM, C.; BORGATTO, A. Um Modelo de Avaliação de Jogos Educacionais na Engenharia de Software. *25th Brazilian Symposium on Software Engineering (SBES)*, 2011.

[4] DAVIS, F.; BAGOZZI, R.; WARSHAW, P. User acceptance of computer technology: A comparison of two theoretical model. management science. *Management Science*, INFORMS, v. 35, n. 8, p. 982–1003, 1989.

[5] COOMANS, S.; LACERDA, G. PETESE, a pedagogical ergonomic tool for educational software evaluation. *6th International Conference on Applied Human Factors and Ergonomics (AHFE) and the Affiliated Conferences.*, Elsevier, v. 3, p. 5881–5888, 2015.

[6] SEYMOUR, E. et al. Using real-world questions to promote active learning. In: ACS. *National Meeting of the American Chemical Society*. San Francisco, 2000.

[7] ROBERTO, G. et al. Tupy Online - Programação em Português com Visualização de Execução e Abstrações de Estruturas de Dados na Web. *26º Workshop sobre Educação em Computação (WEI)*, 2018.

[8] BROOKE, J. SUS- A quick and dirty usability scale. *Usability evaluation in industry*, London: Taylor and Francis, v. 189, n. 194, p. 4–7, 1996.

[9] LEWIS, J. R. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, Taylor & Francis, v. 7, n. 1, p. 57–78, 1995.

[10] ELLSON, J. et al. Graphviz and dynagraph – static and dynamic graph drawing tools. In: *GRAPH DRAWING SOFTWARE*. Heidelberg: Springer-Verlag, 2003. p. 127–148.

[11] GANSNER, E.; KOUTSOFIOS, E.; NORTH, S. Drawing graphs with dot. Technical report, AT&T Research. URL http://www.graphviz.org/pdf/dotguide.pdf, 2015.

[12] PARR, T. *The definitive ANTLR 4 reference*. Raleigh NC: Pragmatic Bookshelf, 2013.

[13] KIRKPATRICK, D.; KIRKPATRICK, J. *Evaluating Training Programs - The Four Levels*. San Francisco: Berrett-Koehler Publishers, Inc, 1994.

[14] KELLER, J. M. Development and use of the arcs model of instructional design. *Journal of instructional development*, Springer, v. 10, n. 3, p. 2, 1987.

[15] BLOOM, B. Taxonomy of educational objectives. vol. 1: Cognitive domain. *New York: McKay*, p. 20–24, 1956.

[16] MOODY, D.; SINDRE, G. Evaluating the Effectiveness of Learning Interventions: An Information Systems Case Study. *European Conference on Information Systems (ECIS)*, p. 80, 2003.

[17] KRATHWOHL, D. A Revision of Bloom's Taxonomy: An Overview. *Theory into practice*, Taylor & Francis, v. 41, n. 4, p. 212–218, 2002.

[18] SNARE, C. An alternative end-of-semester questionnaire. *PS: Political Science & Politics*, Cambridge University Press, v. 33, n. 4, p. 823–825, 2000.

[19] PLAZA, I. et al. Proposal of a quality model for educational software. *20th Annual Conference European Association for Education in Electrical and Information Engineering (EAEEIE)*, 2009.

[20] DEVELLIS, R. *Scale development: Theory and application*. Thousand Oaks, CA, US: Sage Publications, 1991.

[21] CRONBACH, L. Coefficient alpha and the internal structure of tests. *Psychometrika*, Springer, v. 16, n. 3, p. 297–334, 1951.

[22] CASELLA, G.; BERGER, R. *Inferência Estatística*. São Paulo: Cengage learning, 2010.

[23] SAVI, R.; WANGENHEIM, C.; BORGATTO, A. Análise de um modelo de avaliação de jogos educacionais. *Disponível: https://sites.google.com/site/savisites/avaliacao-de-jogos-educacionais. Acessado em: 10 fev. 2019*, 2011.

[24] PETRI, G.; WANGENHEIM, C. How games for computing education are evaluated? a systematic literature review. *Computers & Education*, Elsevier, v. 107, p. 68–90, 2017.

[25] FERREIRA, A. MA-AVA: Modelo de Avaliação da Aprendizagem em Ambientes Virtuais. *Dissertação de Mestrado. Universidade do Estado do Rio de Janeiro*, 2017.

[26] COYNE, L. et al. Exploring virtual reality as a platform for distance team-based learning. *Currents in Pharmacy Teaching and Learning*, Elsevier, v. 10, n. 10, p. 1384–1390, 2018.

[27] CHAUDY, Y.; CONNOLLY, T. Specification and evaluation of an assessment engine for educational games: Empowering educators with an assessment editor and a learning analytics dashboard. *Entertainment Computing*, Elsevier, v. 27, p. 209–224, 2018.

[28] BRONACK, S. et al. Presence pedagogy: Teaching and learning in a 3d virtual immersive world. *International Journal of Teaching and Learning in Higher Education*, ERIC, v. 20, n. 1, p. 59–69, 2008.

[29] DOUMANIS, I. et al. The impact of multimodal collaborative virtual environments on learning: A gamified online debate. *Computers & Education*, Elsevier, v. 130, p. 121–138, 2019.

[30] LEWIS, J. Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies. *International Journal of Human-Computer Interaction*, Taylor & Francis, v. 14, n. 3-4, p. 463–488, 2002.

[31] ADAMS, R. Decision and stress: cognition and e-accessibility in the information workplace. *Universal Access in the Information Society*, Springer, v. 5, n. 4, p. 363–379, 2007.

[32] FRIJTERS, S.; DAM, G.; RIJLAARSDAM, G. Effects of dialogic learning on value-loaded critical thinking. *Learning and Instruction*, Elsevier, v. 18, n. 1, p. 66–82, 2008.

[33] ALFRED, M.; NEYENS, D.; GRAMOPADHYE, A. Learning in simulated environments: An assessment of 4-week retention outcomes. *Applied Ergonomics*, Elsevier, v. 74, p. 107–117, 2019.

[34] MONTGOMERY, D. *Design and analysis of experiments*. Hoboken, NJ: John Wiley & Sons, 2017.

[35] CYBIS, W.; BETIOL, A. H.; FAUST, R. *Ergonomia e usabilidade: conhecimentos, métodos e aplicações*. São Paulo: Novatec editora, 2017.

[36] DUMAS, J.; REDISH, J. *A practical guide to usability testing*. Portland, OR, US: Intellect books, 1999.

[37] RUBIN, J.; CHISNELL, D. *Handbook of usability testing: how to plan, design and conduct effective tests*. Hoboken, NJ: John Wiley & Sons, 2008.

[38] NIELSEN, J.; MACK, R. L. et al. *Usability inspection methods*. New York: Wiley, 1994. v. 1.

[39] LEWIS, J. R. Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In: SAGE PUBLICATIONS. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Los Angeles, 1992. v. 36, n. 16, p. 1259–1260.

[40] LEWIS, J. R.; HENRY, S. C.; MACK, R. L. Integrated office software benchmarks: A case study. In: *Interact*. YorkTown Heights, NY: [s.n.]. p. 337–343.

[41] LEWIS, J. User satisfaction questionnaires for usability studies: 1991 manual of directions for the ASQ and PSSUQ. In: *Tech. Rep. No. 54.609*. Boca Haton: International Business Machines Corporation, 1991.

[42] SAURO, J. Measuring usability with the system usability scale (sus). 2011. *URL: http://www.measuringusability.com/sus.php [accessed 2019-09-25]*, 2011.

[43] SIEGEL, S. *Estatística Não-Paramétrica para as Ciências do Comportamento*. 1º edição. ed. [S.l.]: McGraw-Hill, 1975.

[44] BUSSAB, W.; MORETTIN, P. *Estatística Básica*. 8º edição. ed. São Paulo: Editora Saraiva, 2013.

[45] LEOTTI, V.; BIRCK, A.; RIBOLDI, J. Comparação dos testes de aderência à normalidade kolmogorov-smirnov, anderson-darling, cramer–von mises e shapiro-wilk por simulação. *XI Simpósio de Estatística Aplicada à Experimentação Agronômica*, p. 192, 2005.

[46] SHAPIRO, S.; WILK, B. An analysis of variance test for normality (complete samples). *Biometrika*, JSTOR, v. 52, n. 3/4, p. 591–611, 1965.

[47] SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. 1º. ed. New York: McGraw-Hill, 1956.

[48] GOULART, J.; OLIVEIRA, F.; PINTO, P. Avaliação Sistemática de Eficácia na Aprendizagem de Algoritmos com o uso do TuPy Online. In: SBC. *Anais do XXVII Workshop sobre Educação em Computação*. Belém do Pará, 2019. p. 218–232.

[49] GOULART, J. et al. TuPy Online: Uma Ferramenta para Visualização de Algoritmos. In: CBIE. *V Workshop de Ensino em Pensamento Computacional, Algoritmos e Programação 2019 (WAlgProg)*. Brasília, 2019.

APÊNDICE A – Termo de Consentimento de Participação

OBJETIVO DO ESTUDO

O Estudo tem como objetivo avaliar a influência do TuPy na aprendizagem do aluno na Graduação em Ciência da Computação, além de obter um feedback sobre a usabilidade do programa.

GARANTIA DE SIGILO E PRIVACIDADE

As informações relacionadas ao estudo (notas de prova, este questionário bem como outros julgados necessários) são confidenciais e qualquer informação divulgada em relatório ou publicação será feita sob forma codificada, para que a confidencialidade seja mantida. O pesquisador garante que seu nome não será divulgado sob hipótese alguma.

Diante do exposto acima eu, _____, declaro que fui esclarecido sobre os objetivos do presente estudo. Participo de livre e espontânea vontade do estudo em questão. Foi-me assegurado o direito de abandonar o estudo a qualquer momento, se eu assim o desejar. Declaro também não possuir nenhum grau de dependência profissional ou educacional com os pesquisadores envolvidos nesse projeto (ou seja, os pesquisadores desse projeto não podem me prejudicar de modo algum no trabalho ou nos estudos), não me sentindo pressionado de nenhum modo a participar dessa pesquisa.

Rio de Janeiro, _____ de dezembro de 2018.

_____

Assinatura

APÊNDICE B Questionário de Caracterização

Nome do Participante:_____
Idade:_____
Matrícula:_____
Cr:_____

Sexo:
( )Feminino ( )Masculino

Já cursou esta disciplina antes?
( )Sim ( )Não

Tem familiaridade em Programação?
( )Sim ( )Não

Você trabalha ou estagia?
( )Sim ( )Não

Por quantas horas semanais se dedica ao estudo extraclasse?
( )Normalmente em véspera de prova
( )Até 5 horas
( )Até 10 horas
( )Até 15 horas
( )Até 20 horas ou mais

Quantas disciplinas está cursando neste período?
( )só essa
( )duas
( )três
( )quatro
( )cinco ou mais

Você conhece ferramentas de ensino de programação? Se sim, quais?

_____

APÊNDICE C – Questionário de Percepção

Questionário de Percepção sobre o TUPY

Nome:_____

a) Qual o tempo aproximado de uso do Tupy em casa antes da segunda P1?
_____ horas

b) Por favor, de acordo com a legenda abaixo, circule um número para cada afirmação:

0 – DISCORDO TOTALMENTE 1 – DISCORDO PARCIALMENTE 2 – NÃO DISCORDO, NEM CONCORDO 3 – CONCORDO PARCIALMENTE 4 – CONCORDO TOTALMENTE

Afirmações:

O design do Tupy é atraente	0 1 2 3 4
O Tupy me pareceu interessante desde a apresentação inicial	0 1 2 3 4
A variação (forma ou de atividades) ajudou a me manter atento ao Tupy	0 1 2 3 4
O conteúdo do Tupy é relevante aos meus interesses	0 1 2 3 4
O funcionamento do Tupy está adequado para o aprendizado	0 1 2 3 4
O conteúdo do Tupy está conectado com outros conhecimentos que eu já possuía	0 1 2 3 4
Foi fácil entender o Tupy e começar a utilizá-lo como material de estudo	0 1 2 3 4
Estou satisfeito porque sei que terei oportunidades de utilizar na prática coisas que aprendi com o Tupy	0 1 2 3 4
O Tupy é adequadamente desafiador para mim, não é nem muito fácil nem muito difícil	0 1 2 3 4
Me diverti com o Tupy	0 1 2 3 4
Eu recomendaria o Tupy para meus colegas	0 1 2 3 4
Voltei a utilizar o Tupy outras vezes	0 1 2 3 4
Consegui atingir objetivos no Tupy por meio das minhas habilidades	0 1 2 3 4
Tive sentimentos positivos de eficiência no aprendizado de algoritmos	0 1 2 3 4
Os comandos para realizar ações no Tupy responderam bem	0 1 2 3 4
É fácil aprender a usar a interface e comandos do Tupy	0 1 2 3 4
O Tupy contribuiu para a minha aprendizagem na disciplina	0 1 2 3 4
O Tupy foi eficiente para minha aprendizagem em comparação com outras atividades da disciplina	0 1 2 3 4
A experiência com o Tupy vai contribuir para meu desempenho na vida profissional	0 1 2 3 4

c) Atribua uma nota de 1 a 4 para o seu nível de conhecimento **antes** e **depois** da utilização do Tupy aos conceitos listados abaixo:

1 – FRACO 2 – REGULAR 3 – BOM 4 – ÓTIMO

Conceitos	Lembrar o que é		Compreender como funciona		Aplicar na prática	
	Antes	Depois	Antes	Depois	Antes	Depois
Meu entendimento de Listas Lineares						
Meu entendimento de Filas						
Meu entendimento de Pilhas						

APÊNDICE D – Valores Críticos para Wilcoxon Pareado

N	Nível de significância para prova unilateral		
	0,025	0,01	0,005
	Nível de significância para prova bilateral		
	0,05	0,02	0,01
6	0	-	-
7	2	0	-
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

APÊNDICE E – Valores Críticos para U de Mann-Whitney

Valores críticos de U para uma prova unilateral com alfa = 0,025 e com uma prova bilateral com alfa =0,05

n1 \ n2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2	0	0	0	1	1	1	1	1	2	2	2	2
3	2	3	3	4	4	5	5	6	6	7	7	8
4	4	5	6	7	8	9	10	11	11	12	13	13
5	7	8	9	11	12	13	14	15	17	18	19	20
6	10	11	13	14	16	17	19	21	22	24	25	27
7	12	14	16	18	20	22	24	26	28	30	32	34
8	15	17	19	22	24	26	29	31	34	36	38	41
9	17	20	23	26	28	31	34	37	39	42	45	48
10	20	23	26	29	33	36	398	42	45	48	52	55
11	23	26	30	33	37	40	44	47	51	55	58	62
12	26	29	33	37	41	45	49	53	57	61	65	69
13	28	33	37	41	45	50	54	59	63	67	72	76
14	31	36	40	45	50	55	59	64	67	74	78	83
15	34	39	44	49	54	59	64	70	75	80	85	90
16	37	42	47	53	59	64	70	75	81	86	92	98
17	39	45	51	57	63	67	75	81	87	93	99	105
18	42	48	55	61	67	74	80	86	93	99	106	112
19	45	52	58	65	72	78	85	92	99	106	113	119
20	48	55	62	69	76	83	90	98	105	112	119	127

APÊNDICE F – Questionário SUS e Questionário de *feedback*

Afirmações sobre aspectos gerais:

Eu acredito que gostaria de usar esse sistema com frequência	0 1 2 3 4 NA
Eu acho o sistema desnecessariamente complexo	0 1 2 3 4 NA
Eu achei o sistema fácil de usar	0 1 2 3 4 NA
Eu acho que precisaria de ajuda técnica para ser capaz de usar esse sistema	0 1 2 3 4 NA
Eu achei várias funções do sistema bem integradas	0 1 2 3 4 NA
Eu acho que há muitas inconsistências nesse sistema	0 1 2 3 4 NA
Eu acredito que muitas pessoas podem aprender a usar esse sistema facilmente	0 1 2 3 4 NA
Eu achei o sistema muito complicado de usar	0 1 2 3 4 NA
Eu me senti muito confiante usando o sistema	0 1 2 3 4 NA
Eu preciso aprender muitas coisas antes de poder usar esse sistema	0 1 2 3 4 NA

b) Afirmações sobre aspectos da linguagem TuPy:

É bom que a linguagem básica seja português	0 1 2 3 4 NA
As palavras chaves da linguagem precisam ser melhoradas	0 1 2 3 4 NA
Acompanhar os valores das variáveis é essencial no entendimento do algoritmo	0 1 2 3 4 NA
O esquema de cores para o código precisa ser melhorado	0 1 2 3 4 NA
O completamento de código está adequado	0 1 2 3 4 NA
Não consegui escrever alguns programas por falta de recursos disponíveis	0 1 2 3 4 NA
O esquema de condensação/expansão do editor de códigos é bom	0 1 2 3 4 NA
As Referências Rápidas não ajudam muito a aprender a linguagem	0 1 2 3 4 NA
Os exemplos de algoritmos disponíveis na página ajudam a compreender o TuPy	0 1 2 3 4 NA
Erros de compilação são mal explicados	0 1 2 3 4 NA

Entendi como criar visualizações customizadas	0 1 2 3 4 NA
É difícil criar visualizações customizadas	0 1 2 3 4 NA
A visualização de estruturas de dados ajuda a entender os algoritmos	0 1 2 3 4 NA
É difícil entender como se deve acompanhar a execução de programas	0 1 2 3 4 NA
É importante acompanhar passo a passo a execução de programas	0 1 2 3 4 NA
Falhei em acompanhar execuções passo a passo de programas	0 1 2 3 4 NA
É importante poder ir para frente ou para trás durante a execução de programas	0 1 2 3 4 NA
Erros durante a execução são mal explicados	0 1 2 3 4 NA

c) Comente sobre aspectos da linguagem TuPy:

d) Afirmações sobre a Interface Tupy:

Os componentes estão bem distribuídos na tela	0 1 2 3 4 NA
A visualização de estruturas de dados é clara	0 1 2 3 4 NA
O esquema experimental está bem feito	0 1 2 3 4 NA
Os botões de navegação na execução estão adequados	0 1 2 3 4 NA
A navegação na execução com as setas está adequada	0 1 2 3 4 NA
A exportação de visualização é adequada	0 1 2 3 4 NA
A cópia de código formatado está adequada	0 1 2 3 4 NA
A exportação de elementos está adequada	0 1 2 3 4 NA
O quadro para impressão de saídas está adequado	0 1 2 3 4 NA

As teclas de atalhos para o controle de execução dos programas estão adequadas	0 1 2 3 4 NA
As informações de rodapé são importantes	0 1 2 3 4 NA

e) Comente sobre a interface do TuPy:

f) Afirmações sobre apresentação introdutória do TuPy:

O processo de introdução do TuPy foi adequado	0 1 2 3 4 NA
O suporte ao TuPy foi adequado para mim	0 1 2 3 4 NA
As apresentações orais sobre o TuPy foram esclarecedoras	0 1 2 3 4 NA
As provas P1 foram bem elaboradas para mensurar o quão efetivo foi estudar com o TuPy	0 1 2 3 4 NA
Foi fácil instalar o TuPy em casa	0 1 2 3 4 NA
O estudo dos exemplos em casa foi essencial para mim	0 1 2 3 4 NA
Os exemplos apresentados pelo professor ajudaram a entender	0 1 2 3 4 NA
O número de exemplos foi insuficiente	0 1 2 3 4 NA
Os exercícios pedidos em TuPy foram adequados	0 1 2 3 4 NA
O TuPy deve ser utilizado em outras disciplinas	0 1 2 3 4 NA

g) Comente sobre o processo de introdução do TuPy:

h) Marque quais programas abaixo você simulou:
( )Funções de ListaLinear; ( )Funções de Pilha; ( )Funções de Fila; ( )Notação Polonesa Reversa;
( )Inversão de Pilha; ( )Inversão de Filha; ( )BubbleSort; ( )InsertionSort; ( )MergeSort;
( )QuickSort; ( )SelectionSort; ( )Fatorial recursivo; ( )Fibonacci recursivo;
( )BuscaBinária recursiva;

APÊNDICE G – Valores de a para o teste de Shapiro-Wilk

i\n	2	3	4	5	6	7	8	9	10	11	12	13	
1	0,7071	0,7071	0,6872	0,6646	0,6431	0,6233	0,6062	0,5888	0,5739	0,5601	0,5475	0,5359	
2			0,1677	0,2413	0,2806	0,3031	0,3164	0,3244	0,3291	0,3315	0,3325	0,3325	
3					0,0875	0,1401	0,1743	0,1976	0,2141	0,2260	0,2347	0,2412	
4							0,0561	0,0947	0,1224	0,1429	0,1586	0,1707	
5									0,0399	0,0695	0,0922	0,1099	
6											0,0303	0,0539	
i\n	14	15	16	17	18	19	20	21	22	23	24	25	
1	0,5251	0,5150	0,5056	0,4968	0,4886	0,4808	0,4734	0,4643	0,4590	0,4542	0,4493	0,4450	
2	0,3318	0,3306	0,3290	0,3273	0,3253	0,3232	0,3211	0,3185	0,3156	0,3126	0,3098	0,3069	
3	0,2460	0,2495	0,2521	0,2540	0,2553	0,2561	0,2565	0,2578	0,2571	0,2563	0,2554	0,2543	
4	0,1802	0,1878	0,1939	0,1988	0,2027	0,2059	0,2085	0,2119	0,2131	0,2139	0,2145	0,2148	
5	0,1240	0,1353	0,1447	0,1524	0,1587	0,1641	0,1686	0,1736	0,1764	0,1787	0,1807	0,1822	
6	0,0727	0,0880	0,1005	0,1109	0,1197	0,1271	0,1334	0,1399	0,1443	0,1480	0,1512	0,1539	
7	0,0240	0,0433	0,0593	0,0725	0,0837	0,0932	0,1013	0,1092	0,115	0,1201	0,1245	0,1283	
8			0,0196	0,0359	0,0496	0,0612	0,0711	0,0804	0,0878	0,0941	0,0997	0,1046	
9					0,0163	0,0303	0,0422	0,0530	0,0618	0,0696	0,0764	0,0823	
10							0,0140	0,0263	0,0368	0,0459	0,0539	0,061	
11									0,0122	0,0228	0,0321	0,0403	
12											0,0107	0,0200	
13												0,0000	

ï\n	26	27	28	29	30	31	32	33	34	35	36	37	
1	0,4407	0,4366	0,4328	0,4291	0,4254	0,4220	0,4188	0,4156	0,4127	0,4096	0,4068	0,4040	
2	0,3043	0,3018	0,2992	0,2968	0,2944	0,2921	0,2898	0,2876	0,2854	0,2834	0,2813	0,2794	
3	0,2533	0,2522	0,2510	0,2499	0,2487	0,2475	0,2463	0,2451	0,2439	0,2427	0,2415	0,2403	
4	0,2151	0,2152	0,2151	0,2150	0,2148	0,2145	0,2141	0,2137	0,2132	0,1227	0,2121	0,2116	
5	0,1836	0,1848	0,1857	0,1864	0,1870	0,1874	0,1878	0,1880	0,1882	0,1883	0,1883	0,1883	
6	0,1563	0,1584	0,1601	0,1616	0,1630	0,1641	0,1651	0,1660	0,1667	0,1673	0,1678	0,1683	
7	0,1316	0,1346	0,1372	0,1395	0,1415	0,1433	0,1449	0,1463	0,1475	0,1487	0,1496	0,1505	
8	0,1089	0,1128	0,1162	0,1192	0,1219	0,1243	0,1265	0,1284	0,1301	0,1317	0,1331	0,1344	
9	0,0876	0,0923	0,0965	0,1002	0,1036	0,1066	0,1093	0,1118	0,1140	0,1160	0,1179	0,1196	
10	0,0672	0,0728	0,0778	0,0822	0,0862	0,0899	0,0931	0,0961	0,0988	0,1013	0,1036	0,1056	
11	0,0476	0,0540	0,0598	0,065	0,0697	0,0739	0,0777	0,0812	0,0844	0,0873	0,0900	0,0924	
12	0,0284	0,0358	0,0424	0,0483	0,0537	0,0585	0,0629	0,0669	0,0706	0,0739	0,0770	0,0798	
13	0,0094	0,0178	0,0253	0,032	0,0381	0,0435	0,0485	0,0530	0,0572	0,0610	0,0645	0,0677	
14		0,0000	0,0084	0,0159	0,0227	0,0289	0,0344	0,0395	0,0441	0,0484	0,0523	0,0559	
15				0	0,0076	0,0144	0,0206	0,0262	0,0314	0,0361	0,0404	0,0444	
16						0,0000	0,0068	0,0131	0,0187	0,0239	0,0287	0,0331	
17								0,0000	0,0062	0,0119	0,0172	0,0220	
18										0,0000	0,0057	0,0110	
19												0,0000	

ï\n	38	39	40	41	42	43	44	45	46	47	48	49	50
1	0,4015	0,3989	0,3964	0,3940	0,3917	0,3894	0,3872	0,3850	0,3830	0,3808	0,3789	0,3770	0,3751
2	0,2774	0,2755	0,2737	0,2719	0,2701	0,2684	0,2667	0,2651	0,2635	0,2620	0,2604	0,2589	0,2574
3	0,2391	0,2380	0,2368	0,2357	0,2345	0,2334	0,2323	0,2313	0,2302	0,2291	0,2281	0,2271	0,2260
4	0,2110	0,2104	0,2098	0,2091	0,2085	0,2078	0,2072	0,2065	0,2058	0,2052	0,2045	0,2038	0,2032
5	0,1881	0,1880	0,1878	0,1876	0,1874	0,1871	0,1868	0,1865	0,1862	0,1859	0,1855	0,1851	0,1847
6	0,1686	0,1689	0,1691	0,1693	0,1694	0,1695	0,1695	0,1695	0,1695	0,1695	0,1693	0,1692	0,1691
7	0,1513	0,1520	0,1526	0,1531	0,1535	0,1539	0,1542	0,1545	0,1548	0,1550	0,1551	0,1553	0,1554
8	0,1356	0,1366	0,1376	0,1384	0,1392	0,1398	0,1405	0,1410	0,1415	0,1420	0,1423	0,1427	0,1430
9	0,1211	0,1225	0,1237	0,1249	0,1259	0,1269	0,1278	0,1286	0,1293	0,1300	0,1306	0,1312	0,1317
10	0,1075	0,1092	0,1108	0,1123	0,1136	0,1149	0,1160	0,1170	0,1180	0,1189	0,1197	0,1205	0,1212
11	0,0947	0,0967	0,0986	0,1004	0,1020	0,1035	0,1049	0,1062	0,1073	0,1085	0,1095	0,1105	0,1113
12	0,0824	0,0848	0,0870	0,0891	0,0909	0,0927	0,0943	0,0959	0,0972	0,0986	0,0998	0,1010	0,1020
13	0,0706	0,0733	0,0759	0,0782	0,0804	0,0824	0,0842	0,0860	0,0876	0,0892	0,0906	0,0919	0,0932
14	0,0592	0,0622	0,0651	0,0677	0,0701	0,0724	0,0745	0,0765	0,0783	0,0801	0,0817	0,0832	0,0846
15	0,0481	0,0515	0,0546	0,0575	0,0602	0,0628	0,0651	0,0673	0,0694	0,0713	0,0731	0,0748	0,0764
16	0,0372	0,0409	0,0444	0,0476	0,0506	0,0534	0,0560	0,0584	0,0607	0,0628	0,0648	0,0667	0,0685
17	0,0264	0,0305	0,0343	0,0379	0,0411	0,0442	0,0471	0,0497	0,0522	0,0546	0,0568	0,0588	0,0608
18	0,0158	0,0203	0,0244	0,0283	0,0318	0,0352	0,0383	0,0412	0,0439	0,0465	0,0489	0,0511	0,0532
19	0,0053	0,0101	0,0146	0,0188	0,0227	0,0263	0,0296	0,0328	0,0357	0,0385	0,0411	0,0436	0,0459
20		0,0000	0,0049	0,0094	0,0136	0,0175	0,0211	0,0245	0,0277	0,0307	0,0335	0,0361	0,0386
21				0,0000	0,0045	0,0087	0,0126	0,0163	0,0197	0,0229	0,0259	0,0288	0,0314
22						0,0000	0,0042	0,0081	0,0118	0,0153	0,0185	0,0215	0,0244
23								0,0000	0,0039	0,0076	0,0111	0,0143	0,0174
24										0,0000	0,0037	0,0071	0,0104
25												0,0000	0,0350