

Avaliação Sistemática de Eficácia na Aprendizagem de Algoritmos com o uso do TuPy Online

Jorge L. J. Goulart¹, Fabiano S. Oliveira¹, Paulo E. D. Pinto¹

¹Instituto de Matemática e Estatística
Universidade do Estado do Rio de Janeiro (UERJ)
Rio de Janeiro - RJ - Brazil

jorge.goulart@outlook.com,

{fabiano.oliveira, pauloedp}@ime.uerj.br

Abstract. *This study aims to evaluate 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.*

Resumo. *Este estudo tem como objetivo avaliar 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 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, se considerado o aspecto de percepção dos alunos, um impacto ainda mais expressivo. Ambas análises foram validadas com significância estatística.*

1. 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. [Arruda 2007].

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 tem buscado suprir essa deficiência concebendo e aprimorando alguns métodos de avaliação tais como: *EGame-Flow* [Fu et al. 2009], o *MEEGA* [Savi et al. 2011b], *TAM* [Davis et al. 1989], *PETESE* [Coomans and Lacerda 2015] que serão descritos melhor na Seção 3.

Em 2018, um grupo de pesquisadores da Universidade do Estado do Rio de Janeiro (UERJ), idealizou a ferramenta TuPy Online [Roberto et al. 2018], para amparar e auxiliar estudantes na aprendizagem de algoritmos, e com três objetivos explícitos: (i) introduzir uma pseudolinguagem reduzindo o número de palavras-chave para construção de algoritmos; (ii) basear-se na língua portuguesa com comandos e identificadores escritos em português, atendendo um público-alvo sem domínio da língua inglesa; (iii) permitir ao usuário uma visualização descomplicada do progresso da estrutura de dados frente à depuração do código linha por linha.

Após o desenvolvimento da ferramenta, ela foi experimentalmente utilizada em algumas aulas. Apesar dos comentários positivos de alunos e professores sobre a utilização do TuPy Online, era desejado uma avaliação estatística sobre a eficácia na aprendizagem, bem como uma sistemática análise de sua usabilidade.

Assim sendo, este trabalho tem como objetivo: 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; expor e divulgar os resultados da análise, bem como identificar possíveis melhorias do *software* e do processo de sua introdução visando a aprendizagem.

O trabalho está organizado da seguinte forma. Na Seção 2 apresentaremos com mais detalhes o TuPy Online. Na Seção 3, apresenta-se uma revisão da literatura de métodos de avaliações de aprendizagem. Na Seção 4, explicita-se a metodologia utilizada para análise da aprendizagem com o uso da ferramenta. Na Seção 5, apresenta-se a aplicação da metodologia e discussão dos resultados. Na Seção 6, conclui-se a análise assinalando pontos positivos e negativos da avaliação, além da apresentação de propostas para trabalhos futuros.

2. A ferramenta TuPy Online

A interface do TuPy Online é totalmente baseada no navegador. Conforme mostra a Figura 1, 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. É possível configurar pontos de parada por meio de um clique na linha de código desejada.



Figura 1. A interface de edição de código do TuPy Online.

Existe a possibilidade de o usuário fornecer dados de entrada 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 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 (\neq), por exemplo, resulta na exibição do símbolo \neq . O objetivo é o de aproximar o código TuPy das notações de pseudocódigo manuscrito. Além disso, foi disponibilizada uma versão que pode ser executada localmente, sem necessidade de conexão com a 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* [Ellson et al. 2004], originalmente criado para visualizações de grafos descritos em DOT [Gansner et al. 2015]. A Figura 2 ilustra esse ponto.

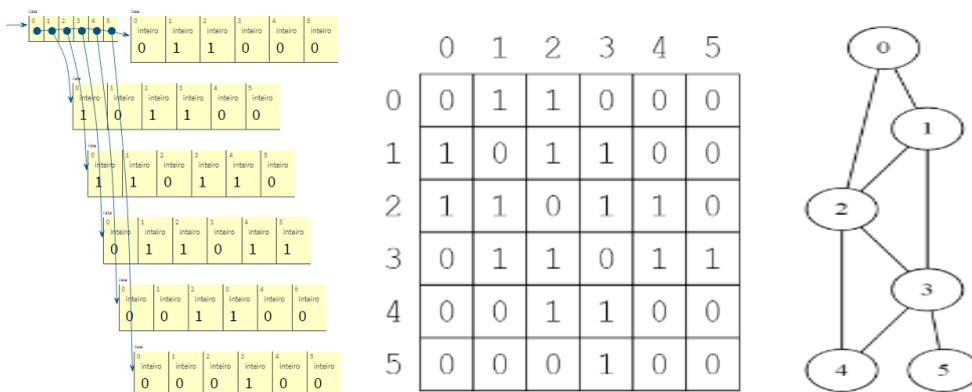


Figura 2. Visualização de uma matriz de adjacência de três formas no TuPy Online: a visualização tradicional do TuPy Online, a visualização com a função de abstração matriz, e com a função de abstração grafo MA.

3. Revisão da Literatura

Nesta seção apresenta-se a revisão da literatura acerca das metodologias empregadas nas avaliações de aprendizagem. 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.

Em [Plaza et al. 2009] 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, 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* [Fu et al. 2009], 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 [DeVellis 1991]. O estudo foi realizado com alunos do curso online “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 [Cronbach 1951], anova e teste-t [Casella and Berger 2010]. Por fim, o questionário foi validado com 42 questões dispersas em 8 fatores.

Em [Savi et al. 2011b], 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 [Kirkpatrick and Kirkpatrick 1994], considerado um padrão da indústria em toda área de recursos humanos e comunidades de treinamento. Esses 4 níveis referem-se a: Reação, Aprendizagem, Comportamento e Resultados. Desses 4 níveis, [Savi et al. 2011b] usam apenas o nível 1, 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 [Keller 1983]. Já o pilar Aprendizagem é estruturado em Conhecimento, Compreensão e Aplicação que são os 3 níveis iniciais da Taxonomia de Bloom [Krathwohl 2002] além da inclusão de 2 outros fatores: Aprendizagem de Longo Termo e Aprendizagem de Curto Termo” [Moody and Sindre 2003]. 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 a 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 [Cronbach 1951]. De acordo com [Savi et al. 2011a], o questionário foi validado com confiabilidade estatística.

Recentemente, a metodologia do *MEEGA* foi adaptada para a elaboração do *MA-AVA* [Ferreira 2017] - 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 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 do aluno. Segundo a autora, o número pequeno de alunos e um curto período de experimentação são limitações da análise.

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 2017] 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.

Considerando os estudos acima, optou-se por adotar como metodologia para o presente trabalho uma adaptação da metodologia de Savi, a inclusão do nível 2 de Kirkpatrick e a utilização de grupos de controle.

4. Metodologia e Aplicação

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), uma turma de Algoritmos e Estrutura de Dados I (AED1) e uma turma de Algoritmo e Estrutura de Dados II (AED2). Além disso, foi utilizado também em 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, inicialmente, um

questionário de caracterização de perfil. Na aula posterior ao preenchimento, 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. 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 3.

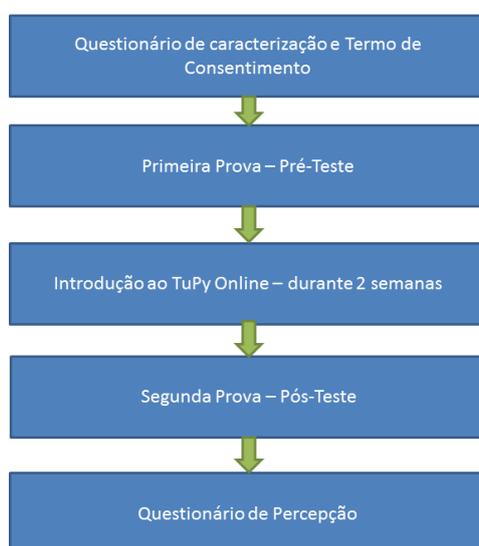


Figura 3. Projeto do experimento.

O grupo de controle foi formulado da seguinte forma: as provas do pré-teste foram divididas em dois modelos *A* e *B* e cada turma dividida em dois grupos de mesmo tamanho (aleatoriamente). O grupo que fez no pré-teste a prova *A*, realizou a prova *B* no pós-teste. Já o outro, realizou o inverso, prova *B* no pré-teste seguida da prova *A* no pós-teste.

Utilizou-se o teste *t* pareado para realizar a análise dos dados. No caso da análise por turmas, cujos tamanhos são pequenos, ficou-se impossibilitado da utilização de ferramentas decisórias baseados em testes paramétricos. Neste caso, utilizaram-se testes não-paramétricos (Wilcoxon pareado e Teste *U* de Mann-Whitney) para as avaliações. Os resultados serão apresentados com o maior nível de confiança possível para aceitação das hipóteses apresentadas.

Por fim, faz-se uma descrição dos questionários aplicados. O questionário de caracterização teve como objetivo o reconhecimento do perfil do participante. As perguntas coletaram dados como: gênero, familiaridade em programação, horas de estudo extra classe, alunos que já cursaram a disciplina, entre outros.

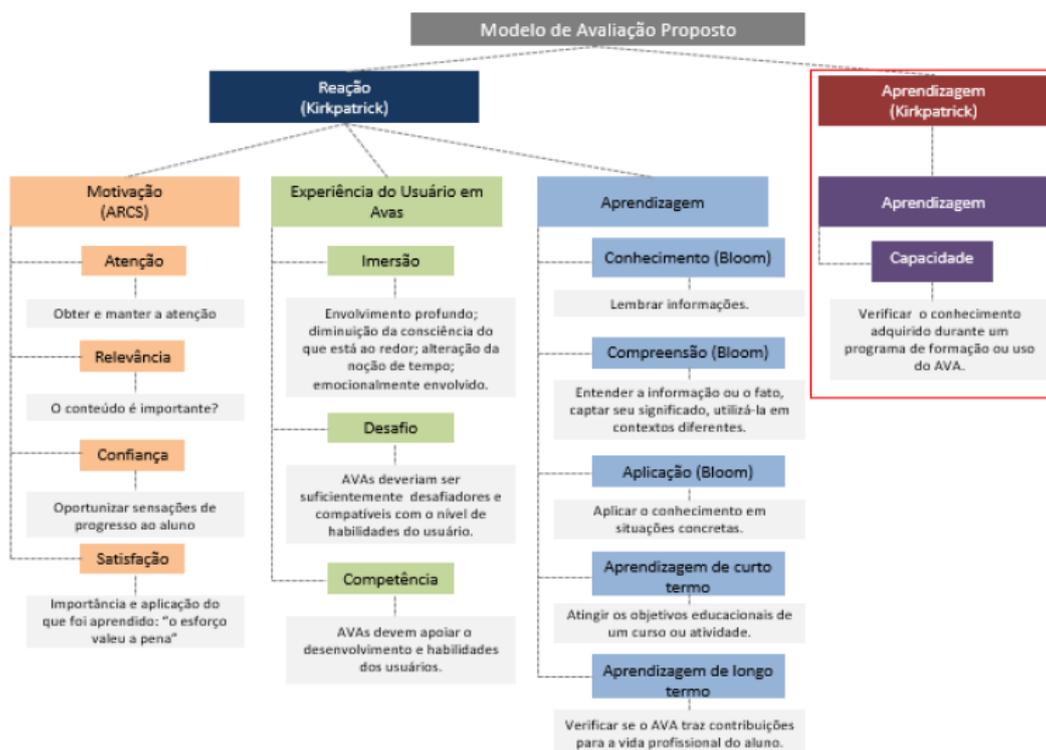


Figura 4. Arcabouço do questionário MA-AVA.

Tendo em vista a adaptabilidade feita por [Ferreira 2017] (Figura 4) do questionário de Savi et al., foram retiradas, para este experimento, as questões consideradas não aplicáveis para um *software* educacional, excluindo a categoria imersão. O questionário resultante para avaliação do TuPy Online é apresentado no Quadro 1. O questionário foi elaborado com 19 questões considerando as dimensões: Atenção, Relevância, Confiança, Desafio, Diversão, Competência e Controle. Além de considerar a Aprendizagem de curto termo e a Aprendizagem de longo termo.

Anexo ao questionário de percepção, o aluno realizou uma auto-avaliação, seguindo os princípios da Taxonomia de Bloom. O estudante indicaria um valor entre 1 e 4 (Fraco, Regular, Bom, Ótimo) para seu conhecimento anterior e posterior a utilização do TuPy Online, dividido em 3 categorias (Lembrança, Compreensão e Aplicabilidade).

5. Apresentação dos Resultados

Nesta seção, apresentam-se os dados, resultados, análises e discussões do experimento realizado. Primeiramente é apresentada a caracterização dos estudantes, obtida com o primeiro questionário, pelas Figuras 5 a 8.

Cabe ressaltar que todos os alunos que participaram do experimento estiveram em contato com o TuPy Online nas 2 semanas de aplicação. No entanto, como a participação era facultativa, alguns alunos que realizaram o pré-teste e pós-teste não quiseram responder o questionário final, bem como alguns que, por algum motivo, não participaram de um dos testes, quiseram responder os questionários e foram considerados válidos.

A Figura 5 mostra a distribuição da turma por gênero. Mais de 80% dos participantes foram do sexo masculino. O baixo número da presença feminina, impossibilita

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 possuía
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	Me diverti 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

Quadro 1. Questionário para avaliação de percepção do TuPy.

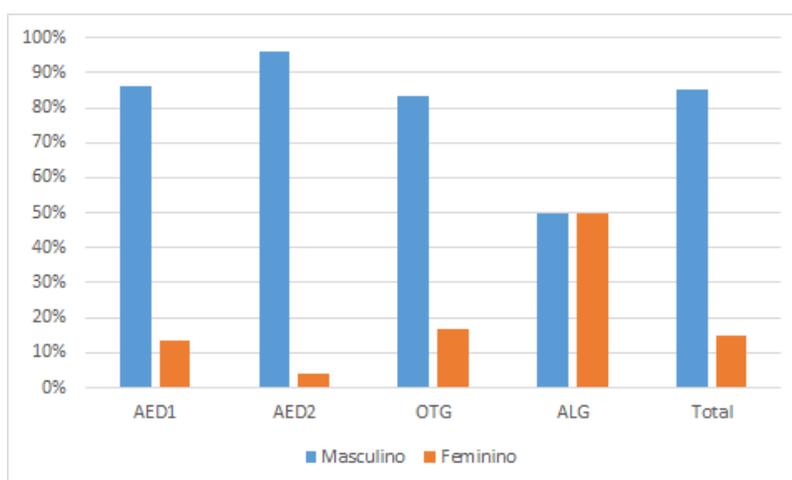


Figura 5. Disposição das turmas por gênero.

análises por gênero.

A Figura 6 representa a distribuição dos alunos que já cursaram a disciplina. Mais de 70% em todas as turmas não cursaram a disciplina anteriormente. Este é um fato importante, pois o aprendizado com o TuPy poderia ser de menor influência em alunos já expostos aos assuntos abordados.

A Figura 7 mostra a familiaridade dos alunos com programação. Nota-se que a frequência de alunos é maior que 70% e se apresentou de maneira uniforme.

A Figura 8, 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.

A seguir, as Figuras 9 e 10, apresenta-se o resultado dos dados obtidos do ques-

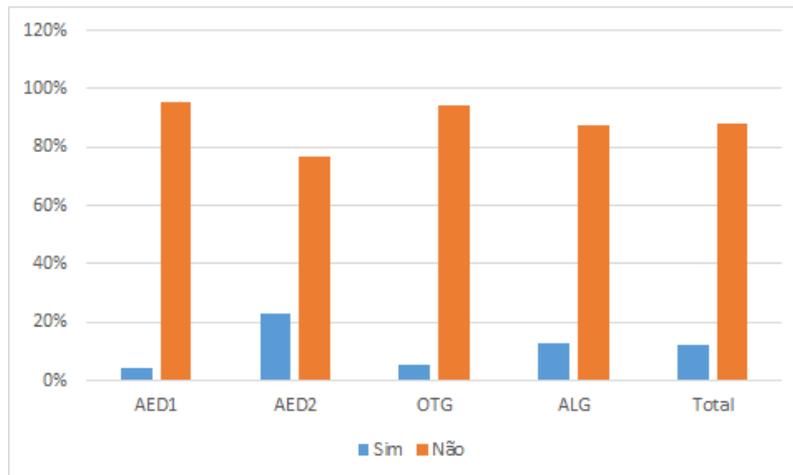


Figura 6. Alunos que já cursaram a disciplina.

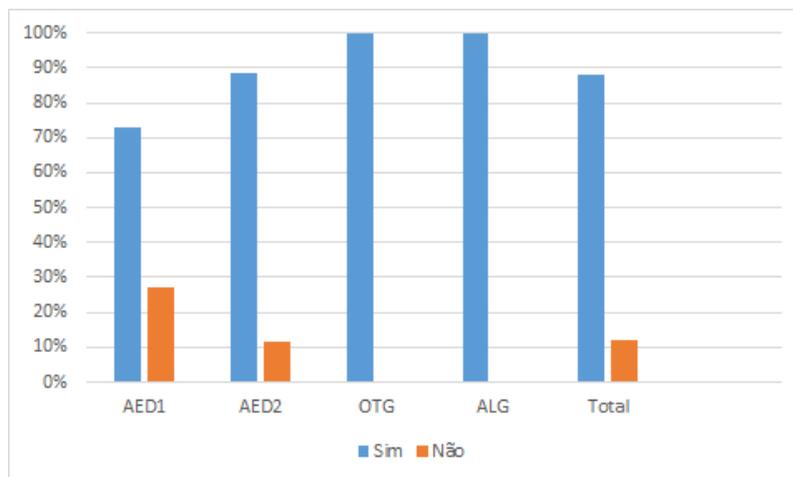


Figura 7. Alunos com familiaridade em programação.

tionário de percepção. A Figura 9 representa as 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 na aprendizagem (Figura

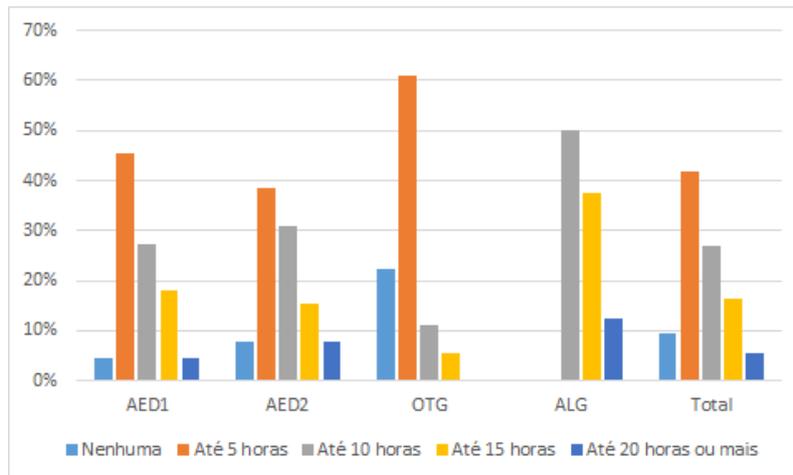


Figura 8. Horas de estudo extra classe.

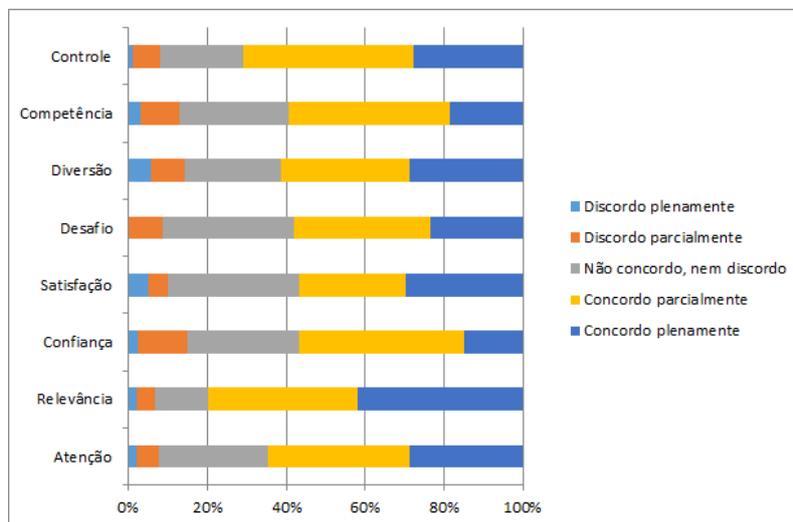


Figura 9. Avaliação do questionário de percepção.

10), 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 os 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.

A Tabela 1 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

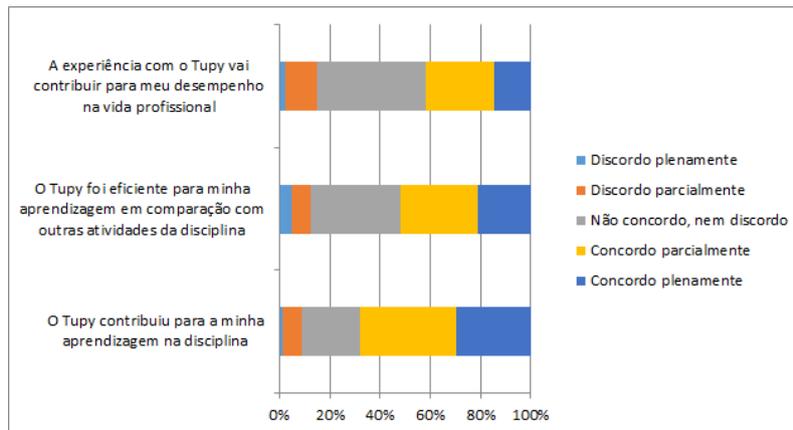


Figura 10. Aprendizagem de curto e longo prazo.

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

Tabela 1. Resultados da avaliação pela Taxonomia de Bloom.

programação dinâmica, não perceberam melhorias após a utilização do TuPy Online.

A Tabela 2 representa os resultados do pré-teste e pós-teste e a Tabela 3 representa as médias e medianas destes testes. Na avaliação de aprendizagem pelas notas obtidas, considerando que as amostras são dependentes, foi realizado o teste de Wilcoxon pareado [Siegel 1975]. Para os grupos de controle (um grupo é controle do outro), foi utilizado o teste U de Mann-Whitney. A hipótese nula adotada é a não influência do TuPy Online. Desta forma, tem-se os resultados a seguir.

Considerando todos os alunos de todas as disciplinas, 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 0,0101. Em outras palavras, houve uma melhora significativa no aprendizado dos alunos com a utilização do TuPy Online.

No caso da disciplina AED1, em que 23 alunos participaram, a análise intragrupo resultou em um valor observado de 37,5 para o Grupo 1 frente ao valor crítico de Wilcoxon (14) 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 re-

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
1	8,3	8,1	1,9	5,3	1	5,0	5,5	6,5	7,0
2	4,9	0,2	1,2	1,0	2	6,5	2,0	7,5	9,5
3	3,2	6,5	7,7	7,3	3	8,3	8,5	4,5	6,0
4	6,1	7,0	5,8	7,1	4	6,5	9,8	2,5	4,0
5	6,8	7,0	5,7	3,2	5	0,5	5,0	4,0	6,0
6	5,9	7,4	2,7	1,2	6	5,0	4,5	4,0	6,8
7	1,7	0,1	4,1	6,7	7	10,0	6,9	6,5	10,0
8	4,0	4,4	0,1	0,1	8	7,5	2,5	4,0	5,5
9	7,5	7,2	9,4	9,0	9	3,5	4,5	8,5	7,5
10	7,3	4,8	7,0	6,3	10	3,5	2,5	3,3	3,4
11	9,0	9,1	2,5	4,3	11	7,0	9,8		
12	0,5	0,9			12	2,0	8,5		
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
1	4,5	4,0	3,9	8,0	1	6,7	7,6	9,9	9,8
2	5,0	4,5	3,0	4,8	2	7,0	8,3	8,1	8,4
3	5,0	7,0	4,5	6,8	3	7,4	8,1	5,2	5,7
4	7,0	5,5	2,0	2,3	4	5,4	9,9	6,2	4,5
5	1,5	0,5	2,0	3,2	5	6,4	7,5	7,7	8,9
6	4,3	3,3	2,3	6,0	6			9,5	9,7
7	4,3	2,8	3,5	7,5	7			7,6	9,6
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							

Tabela 2. Avaliação dos alunos em cada disciplina - Pré e Pós-teste.

sultado análogo para o Grupo 2, com estatística de teste igual a 24 e valor crítico de 11.

Na disciplina de OTG, 22 alunos participaram. O teste de Wilcoxon rejeitou a hipótese nula para o Grupo 2, com valor observado de Wilcoxon de 3 frente ao valor crítico de 8. No entanto, para o Grupo 1, não houve rejeição da hipótese nula, com valor observado de 34,5 e valor crítico 14. Para o Grupo 2, há indicação que o mesmo foi influenciado positivamente pelo TuPy Online.

Na disciplina de AED2 houve participação de 28 alunos, e assim como em OTG, só o Grupo 2 representou melhoria significativa, ou seja, através do teste de Wilcoxon pareado, encontrou-se evidências para rejeição da hipótese nula. No Grupo 1 encontrou-se valor observado de 37 com valor crítico de 21. Para o Grupo 2 o valor observado foi

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
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
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

Tabela 3. Médias e Medianas da avaliação dos alunos em cada disciplina - Pré e Pós-teste.

de 4, 5 com valor crítico de 17. Como o valor observado para o Grupo 2 foi menor que o valor crítico, rejeita-se a hipótese nula sob o nível de 95% de confiança.

Já na disciplina de ALG, o TuPy demonstrou diferença apenas em um grupo, rejeitando a hipótese nula. No Grupo 1 obteve-se valor observado de 0 em comparação ao valor crítico de 0. No Grupo 2, o valor observado foi de 7 com valor crítico de 2. Neste Grupo, rejeita-se a hipótese nula, contudo, se desconsiderarmos os alunos que reportaram não terem estudado em casa com o TuPy Online.

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. Sendo assim, há a possibilidade de que a melhoria observada seja parcialmente devida à provas com níveis de dificuldades distintas.

6. Conclusão

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.

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.

Uma possibilidade para explicar o melhor desempenho dos alunos mestrados, 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 em ciências computacionais 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 programação, 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.

A partir desta experiência na avaliação do TuPy Online para algumas turmas de Computação pode-se indicar alguns outros estudos de utilidade. A aplicação de testes de usabilidade poderiam indicar melhorias para tornar a ferramenta mais atraente e útil para as turmas de Computação. Outros estudos, considerando 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 outras disciplinas da Ciência da Computação, bem como em outros cursos de graduação nos quais programação é utilizada, tais como Estatística, Engenharia e afins.

Referências

- Arruda, J. (2007). *Modelagem do Processo de Aprendizagem na Educação Superior*. edUERJ.
- Casella, G. and Berger, R. (2010). *Inferência Estatística*. Cengage learning.
- Coomans, S. and Lacerda, G. (2015). PETESE, a pedagogical ergonomic tool for educational software evaluation. *6th International Conference on Applied Human Factors and Ergonomics (AHFE) and the Affiliated Conferences*.
- Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.
- Davis, F., Bagozzi, R., and Warshaw, P. (1989). User acceptance of computer technology: A comparison of two theoretical model. *management science*. *Management Science*, 35(8):982–1003.
- DeVellis, R. (1991). *Scale development: Theory and application*. Sage Publications.
- Ellson, J., Gansner, E., Koutsofios, E., North, S., and Woodhull, G. (2004). Graphviz and dynagraph – static and dynamic graph drawing tools. In *Graph drawing software*, pages 127–148. Springer.
- Ferreira, A. (2017). MA-AVA: Modelo de Avaliação da Aprendizagem em Ambientes Virtuais. *Dissertação de Mestrado*. Universidade do Estado do Rio de Janeiro.
- Fu, F., Su, R., and Yu, S. (2009). Egameflow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education*, 52(1):101–112.
- Gansner, E., Koutsofios, E., and North, S. (2015). Drawing graphs with dot.
- Keller, J. (1983). Development and use of the ARCS model of motivational design. Technical report, Exeter, England: at the annual meeting of the Association for Education and Training Technology.
- Kirkpatrick, D. and Kirkpatrick, J. (1994). *Evaluating Training Programs - The Four Levels*. Berrett-Koehler Publishers, Inc.
- Krathwohl, D. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory into practice*, 41(4):212–218.
- Montgomery, D. (2017). *Design and analysis of experiments*. John Wiley & Sons.
- Moody, D. and Sindre, G. (2003). Evaluating the Effectiveness of Learning Interventions: An Information Systems Case Study. *European Conference on Information Systems (ECIS)*.
- Plaza, I., Marcuello, J., R.Igual, and Arcega, F. (2009). Proposal of a quality model for educational software. *Annual Conference European Association for Education in Electrical and Information Engineering (EAEEIE)*.
- Roberto, G., Oliveira, F., Pinto, P., and Coelho, I. (2018). Tupy Online - Programação em Português com Visualização de Execução e Abstrações de Estruturas de Dados na Web. *26º Workshop sobre Educação em Computação (WEI)*.

- Savi, R., Wangenheim, C., and Borgatto, A. (2011a). 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.*
- Savi, R., Wangenheim, C., and Borgatto, A. (2011b). Um Modelo de Avaliação de Jogos Educacionais na Engenharia de Software. *25th Brazilian Symposium on Software Engineering (SBES)*.
- Siegel, S. (1975). *Estatística Não-Paramétrica para as Ciências do Comportamento*. McGraw-Hill.