Ensaio de Proficiência em Avaliação da Conformidade de Produtos de Software

Jailton Santos das Neves[1], Paulo Roberto de M. Nascimento[2], Wladmir Araujo Chapetta[2], Raphael Carlos S. Machado[1,2], Vânia de Oliveira Neves[1], Wilson de Souza Melo Júnior[2], Márcio Eduardo Delamaro[3], Simone do Rocio S. de Souza[3], Paulo Sergio L. de Souza[3], Felipe Diniz Dallilo[3]

[1]Instituto de Computação – Universidade Federal Fluminense (UFF)

Campus da Praia Vermelha – Boa Viagem – 24.210-346 – Niterói – RJ – Brasil

jailton_neves@id.uff.br, vania@ic.uff.br

[2]Instituto Nacional de Metrologia, Qualidade e Tecnologia (Inmetro)

Av. Nossa Senhora das Graças, 50 – Xerém – 25.250-020 – Duque de Caxias – RJ – Brasil

{prnascimento,wachapetta,rcmachado,wsjunior}@inmetro.gov.br

[3]Instituto de Ciências Matemáticas e Computação – Universidade de São Paulo (USP)

Av. Trabalhador São-carlense, 400 – Centro – 13.566-590 – São Carlos – SP – Brasil

{delamaro,srocio,pssouza}@icmp.usp.br, felipedallilo@hotmail.com

Abstract. *This work presents a study on the proficiency of laboratories where their competence in the analysis of cyber products is evaluated. The methods, scenarios and tools used for the evaluation proposed here in Brazil and the evaluation criteria will be presented. The results obtained by the laboratories participating in the test rounds, the challenges and future research work will also be presented.*

Resumo. *Este trabalho apresenta um estudo sobre a proficiência de laboratórios onde são avaliadas suas competências na análise de produtos cibernéticos. Serão apresentados os métodos, cenários e ferramentas utilizadas para a avaliação proposta aqui no Brasil e os critérios de avaliação. Também serão apresentados os resultados obtidos pelos laboratórios participantes das rodadas de teste, os desafios e trabalhos futuros de pesquisa.*

1. Introdução

A avaliação de produtos de software por laboratórios acreditados é uma demanda crescente em todo o mundo [INMETRO, 2019]. Isso ocorre devido à forte tendência de se utilizar software para automação, gerenciamento, controle e otimização dos mais diversos dispositivos. Deste modo, a avaliação de um produto de software constitui um mecanismo imprescindível para se atestar a conformidade de um produto em relação a um conjunto de requisitos funcionais e não funcionais, garantindo deste modo, sua eficiência, confiabilidade e segurança [INMETRO, 2020].

DOI: 10.17648/wrac-wciber-2020-125131
https://proceedings.science/p/125131?lang=en

Embora Ensaios de Proficiência (EP) e comparações interlaboratoriais sejam realizados em diferentes escopos de avaliação da conformidade, a avaliação de software introduz uma série de novos desafios decorrentes do grau de subjetividade associado à interpretação dos requisitos de um produto de software, bem como sobre os diferentes entendimentos a respeito de como um software pode ser testado [Myers, 2011; Machado, 2020]. Dessa forma, avaliar a competência de um laboratório na certificação de um produto de software não é uma tarefa trivial. De fato, uma busca nos documentos dos principais organismos acreditadores internacionais mostra que o problema de avaliar proficiência de laboratórios na área de software é ainda um problema em aberto [INMETRO, 2020]. Ainda assim, dada a importância crescente de tais laboratórios para a realização de avaliações de segurança de software de um número cada vez maior de dispositivos inteligentes e ativos de tecnologia, torna-se premente propor abordagens para iniciar o tratamento de tal problema.

Este artigo descreve os resultados da primeira rodada de Ensaio de Proficiência realizada em 2019 [INMETRO, 2019] e de uma segunda rodada, sob a forma de uma comparação interlaboratorial, denominada Primeira Rodada Pública de Proficiência em Avaliação da Conformidade em Produtos de Software. Desde 2019 vem-se utilizando essas rodadas como comparações interlaboratoriais no escopo relacionado à Avaliação Produto de Software (APS), com o intuito de desenvolver pesquisas relacionadas à medição quantitativa do desempenho de laboratórios e equipes de teste. Assim, estas rodadas visam obter um benchmarking, a partir de desafios propostos por um provedor dessas comparações, como um provedor de Ensaios de Proficiência (EP) oficial ou não, almejando de usá-las para identificar lacunas a serem preenchidas e propor melhorias relacionadas às práticas de Verificação, Validação e Testes (VV&T) executadas por esses laboratórios.

A primeira rodada de comparações interlaboratoriais baseou-se em medidas de cobertura de código como principal elemento de intercomparação objetiva dos resultados obtidos por cada laboratório, onde cada laboratório participante teve a oportunidade de avaliar seus métodos e procedimentos de APS e verificar a harmonia com o desempenho dos outros laboratórios [Machado, 2020].

Para a segunda rodada, foi proposto que a avaliação do desempenho dos participantes (laboratórios, profissionais, estudantes, organizações de software etc.) se baseasse na razão de duas medidas: o percentual de cobertura de mutantes (escore de mutação [Delamaro, 2016] dividido pelo tamanho dos casos de teste. O elemento numerador dessa razão visa expressar, em termos percentuais, e quantificar a qualidade e a variabilidade das entradas dos casos de teste produzidos por uma equipe de teste de software. O elemento denominador visa beneficiar casos de testes projetados mais eficientemente, como uma medida indireta da quantidade de combinações de entradas, ou seja, casos de teste com menos combinações de entradas encontrando os mesmos defeitos são mais eficientes, pois demandam menos esforço e recursos de construção do que aqueles com mais combinações para alcançar escore de mutação iguais. Após a realização da rodada e a análise dos resultados pelas instituições coordenadoras (Inmetro/UFF/ICMC-USP), cada participante também teve a oportunidade de avaliar seus métodos e procedimentos de APS e verificar a harmonia com base na comparação de seu desempenho com os de outros laboratórios.

DOI: 10.17648/wrac-wciber-2020-125131
https://proceedings.science/p/125131?lang=en

Em todas as rodadas foram entregues pelos participantes os casos de testes desenvolvidos e um Relatório de Ensaio, o que permitiu obter não apenas uma análise quantitativa dos resultados, mas também uma análise qualitativa, tomando como base os *feedbacks* fornecidos pelos participantes de cada rodada.

Esse artigo tem como objetivo evidenciar os resultados obtidos na avaliação dessas duas rodadas, que por sua vez visam, desenvolver modelos de avaliação de desempenho dos laboratórios em testes de produtos de software. Dessa forma, esse artigo está organizado da seguinte maneira: a seção 2 explora as métricas e os critérios de avaliação das rodadas. A seção 3 apresenta os resultados dos participantes das rodadas realizadas. Na seção 4 é realizada a análise comparativa das duas rodadas e por fim, na seção 5 são apresentadas as considerações finais.

2. Rodadas de Ensaio de Proficiência

O Inmetro provê Ensaios de Proficiência para comparação interlaboratoriais no escopo de APS, um trabalho pioneiro no Brasil quanto a acreditação de laboratórios e avaliação de produtos de software, mas internacionalmente, algumas instituições já realizam essa tarefa através de *Laboratory Accreditation Programs (LAPs),* como o NIST (*National Institute of Standards and Technology*) dos Estados Unidos [NIST, 2017] e o ANSSI *(Autorité Nationale en matière de Sécurité et de défense des Systèmes d'Information)* da França [ANSSI, 2015]. Aqui, até o momento foram realizadas duas rodadas comparações interlaboratoriais, a primeira foi fechada e somente laboratórios acreditados no escopo de APS participaram, na segunda rodada que na verdade foi a primeira rodada pública, puderam participar laboratórios, estudantes e profissionais que mostraram interesse.

2.1. Primeira Rodada

Na primeira rodada, foi utilizada a cobertura de código, que é uma métrica utilizada em análise de teste de software para identificar a quantidade de linhas do código-fonte que foram testadas [Milano, 2011]. Segundo [Câmara, 2019], a metodologia aplicada nesta rodada consistia de utilizar técnicas de cobertura de código e testes de unidade com um objetivo diferente da abordada em Engenharia de software, onde os testes são utilizados com o intuito de minimizar erros e riscos envolvidos com o processo de desenvolvimento de *software* [Maldonado, 1998]. Na metodologia adotada para essa rodada, é definido que, dado um código-fonte de um software (item de ensaio) e dado um conjunto de linhas cobertas deste código (relatório de referência) por um teste de unidade de referência implementado pelo provedor da rodada, o laboratório participante deveria tentar cobrir o mesmo conjunto de linhas, sem conhecer o teste de unidade de referência com um prazo de, aproximadamente, 1 mês para a entrega dos resultados de suas análises. Esta metodologia por cobertura de código parte do princípio que a capacidade de inspecionar códigos-fonte de um software é uma competência requerida para laboratórios envolvidos com análise de software.

Para que os participantes pudessem analisar o código-fonte do software usado na rodada, o provedor do ensaio gerou casos de testes de unidade e os relatórios de referência, que continham informações sobre as linhas cobertas do código-fonte, porém foram apresentados aos participantes apenas os relatórios de referências e não os testes de unidade implementados. Os laboratórios participantes puderam então, implementar seus próprios testes de unidade e emitir seus relatórios, contendo o conjunto de linhas do código-fonte por eles cobertos. Os laboratórios participantes tinham que entregar tanto

DOI: 10.17648/wrac-wciber-2020-125131
https://proceedings.science/p/125131?lang=en

seus relatórios quanto seus testes de unidades para serem avaliados pelo provedor da rodada [INMETRO, 2019].

O provedor, por sua vez, calculou uma nota em relação à proficiência de cada laboratório de acordo com as métricas de avaliação pré-estabelecidas. Caso o laboratório obtivesse uma boa avaliação, pode-se concluir que ele tem um bom entendimento do código relacionado àqueles testes e, por consequência, que o laboratório é proficiente na atividade de análise de software [Machado, 2020].

2.1.1. Critérios de avaliação da primeira rodada

Uma vez que o objetivo nessa rodada consistia em encontrar casos de teste que cobrissem exatamente as mesmas linhas de código cobertas pelos testes de referência, fez se necessário propor um mecanismo de avaliação que levasse em consideração esse aspecto. Sendo assim, o desempenho do participante seria melhor quando seus testes fossem mais similares aos testes de referência. Então, para um dado caso de teste baseado em cobertura de código, foi denominado **T** o conjunto total de linhas do software avaliado, S_R o conjunto de linhas cobertas pelos testes de referência e S_L o conjunto de linhas cobertas pelos testes executados pelo participante. Como métrica de similaridade usada para medir o desempenho de cada participante foi adotado o Índice de Jaccard [Jaccard, 1912], definido a seguir:

$$J(S_R, S_L) = \frac{S_R \cap S_L}{S_R \cup S_L}$$

A determinação do valor de consenso desta comparação interlaboratorial foi através da média dos índices $J(S_R, S_L)$ obtidos por cada um dos participantes. Assim, o valor $\underline{J}$ é dado por:

$$\underline{J} = \sum_{i=1}^{labs} \frac{J(S_R, S_L^i)}{labs}$$

onde $labs$ é o número de participantes. A interpretação do desempenho do i-ésimo laboratório em relação aos demais participantes foi associada à comparação de seu Índice de Jaccard, com o desvio padrão, dado por:

$$\sigma = \sqrt{\sum_{i=1}^{labs} \frac{(\underline{J} - J(S_R, S_L^i))^2}{labs}}$$

O resultado da avaliação do índice do i-ésimo participante (J_i) foi dado pelas seguintes regras:

$J_i \geq \underline{J} - \sigma$ indica desempenho "satisfatório" e não gera sinal;

$\underline{J} - \sigma > J_i \geq \underline{J} - 3\sigma$ indica desempenho "questionável" e gera um sinal de alerta;

$J_i < \underline{J} - 3\sigma$ indica desempenho "insatisfatório" e gera um sinal de ação.

Os resultados dos índices J_i foram arredondados com duas casas decimais, obedecendo aos critérios de arredondamento.

DOI: 10.17648/wrac-wciber-2020-125131
https://proceedings.science/p/125131?lang=en

2.2. Segunda Rodada

Para a segunda rodada, o método adotado estava relacionado a conceitos de teste de mutação, como apresentado por [Delamaro, 2016], através do qual pretendia-se verificar a qualidade dos testes desenvolvidos pelos participantes. O teste de mutação se baseia na inserção de defeitos no código original para gerar variações que permitam identificar a ausência de entradas e, consequentemente, pontos de falha que os casos de teste desenvolvidos poderiam ou deveriam revelar. As variações do código original são denominadas mutantes e cada mutante contém apenas uma modificação em relação ao código original. A modificação é gerada a partir de um conjunto pré-definido de operadores de mutação [Delamaro, 2016].

O desafio proposto aos participantes é que estes produzam casos de teste que eficientemente eliminem o maior número de mutações possíveis para as classes selecionadas pelo comitê organizador da rodada, dentro do prazo de 4 dias. As classes foram selecionadas considerando a quantidade de mutantes gerados pela ferramenta PIT[1], adotando o conjunto completo de operadores de mutação disponibilizado por ela.

2.2.1. Critérios de avaliação da segunda rodada

A avaliação de desempenho de cada participante foi baseada na relação entre a quantidade de mutantes mortos e o tamanho, em KBytes, do conjunto de testes desenvolvido [INMETRO, 2020]. Cada participante deveria então proceder com a avaliação do software e tentar modelar um ou mais casos de teste com o objetivo de matar o maior número de mutantes possível. Desta forma, os participantes que conseguissem eliminar um número elevado de mutantes com um número reduzido de casos de teste estariam sendo mais eficientes. Partindo desta premissa, seria possível concluir quais participantes teriam um bom entendimento da documentação do software, do código relacionado àqueles testes, de como os testes deveriam ser executados e, por consequência, concluir quais participantes seriam proficientes na execução de um APS.

Uma vez que o objetivo consiste em encontrar casos de teste que eliminem o maior número de mutantes possíveis, fez-se necessário propor um mecanismo de avaliação que levasse em consideração esse aspecto. Para avaliar a adequação de um conjunto de casos de teste em relação ao programa original, foi utilizado o escore de mutação, o qual pode variar de 0% a 100% e é definido pela seguinte equação:

$$escore = \frac{M}{T - E}$$

onde $\boldsymbol{M}$ é o número de mutantes mortos com os casos de testes desenvolvidos, $\boldsymbol{E}$ é o número de mutantes equivalentes e $\boldsymbol{T}$ é o total de mutantes gerados. Como na rodada não se tem registro dos mutantes equivalentes, o cálculo de escore de mutação foi composto apenas por:

$$escore = \frac{M}{T}$$

[1]A ferramenta PIT (https://pitest.org/, acessado em 13/11/2020) foi utilizada através de um plugin para a IDE Eclipse, o que permitia uma análise rápida das quantidades de mutantes gerados e mortos pelos casos de testes desenvolvidos pelos participantes, através de um sumário gerado pelo próprio plugin.

DOI: 10.17648/wrac-wciber-2020-125131
https://proceedings.science/p/125131?lang=en

De acordo com os procedimentos disponíveis para o estabelecimento de valores designados pela ABNT NBR ISO/IEC 17043:2011 [ISO 17043, 2011], os valores designados desta rodada foram calculados através de métodos estatísticos descritos no item 7.7 da norma ISO 13528:2015 [ISO 13528, 2015], ou seja, valores de consenso de participantes.

A determinação do valor de consenso desta rodada de comparação foi através da média (μ) das razões entre escore de mutação e o tamanho dos casos de teste de cada participante, dado por:

$$\mu = \sum_{i=1}^{n} \frac{(S_i/t_i)}{n}$$

onde S_i é o escore de mutação e t_i é o tamanho dos casos de teste do participante i, medido em Kbytes do código objeto (bytecode Java), e n é o número total de participantes.

A interpretação do desempenho do i-ésimo participante em relação aos demais foi associada à comparação de seu índice de desempenho com o desvio padrão, dado por:

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(R_i - \mu)^2}{n-1}}$$

onde $R_i = S_i / t_i$, é o indicador de desempenho individual do participante i. O resultado da avaliação do índice do i-ésimo participante foi dado pelas seguintes regras:

$R_i > \mu + 3\sigma$	indica desempenho “excepcional” e não gera sinal;
$\mu + 3\sigma \geq R_i \geq \mu + 2\sigma$	indica desempenho “muito bom” e não gera sinal;
$\mu + 2\sigma \geq R_i \geq \mu + \sigma$	indica desempenho “bom” e não gera sinal;
$\mu + \sigma \geq R_i \geq \mu - \sigma$	indica desempenho “satisfatório” e não gera sinal;
$\mu - 3\sigma \leq R_i \leq \mu - \sigma$	indica desempenho “aceitável” e gera um sinal de alerta;
$R_i \leq \mu - 3\sigma$	indica desempenho “insatisfatório” e gera um sinal de ação.

Os resultados dos índices foram arredondados com cinco casas decimais, obedecendo aos critérios de arredondamento, devido ao fato dos valores de R_i se revelarem muito baixos.

3. Resultados e desempenho dos participantes das Rodadas

Serão apresentados nessa seção os resultados e o desempenho dos participantes das duas rodadas realizadas, a primeira com ênfase na cobertura de código e a segunda que passou a utilizar o escore de mutação.

3.1. Resultados da primeira rodada

A Tabela 1 apresenta os resultados de cada participante [INMETRO, 2019], medidos pelo índice de Jaccard. Como é possível observar, todos os participantes obtiveram valor máximo, com índice J_i = 1. Tal resultado implica que todos os participantes foram capazes de solucionar o desafio de testes referente a esta rodada de EP, e

DOI: 10.17648/wrac-wciber-2020-125131
https://proceedings.science/p/125131?lang=en

consequentemente implementaram casos de teste que cobrem exatamente o conjunto de linhas de código que constituem o desafio proposto.

Tabela 1 - Resultados da Primeira Rodada. [INMETRO, 2019]

Código do Participante	Índice de Jaccard
01	1,00
05	1,00
07	1,00
12	1,00
14	1,00
18	1,00

Em função dos resultados apresentados na Tabela 1, tem-se que o desvio padrão associado aos valores do índice de Jaccard é zero. Portanto, a aplicação dos critérios estabelecidos indica que todos os participantes apresentaram resultado satisfatório.

O fato de todos os participantes apresentarem resultado idêntico aponta duas questões fundamentais. A primeira é que a natureza do desafio de teste proposto por essa comparação interlaboratorial permite que os participantes repitam seus procedimentos de ensaio até obterem um caso de teste que cobre exatamente o conjunto de linhas de código proposto no desafio. A segunda questão é a evidência de que o desafio de testes proposto não apresentou dificuldade suficiente para os participantes, assim como foi relatado por eles em seus Relatórios de Ensaio.

3.2. Resultados da segunda rodada

Para os cálculos dos escores de mutação de cada participante, foram considerados os 2938 mutantes gerados pela ferramenta PIT para o código-fonte analisado nesta rodada. Como a rodada foi pública, um grupo de alunos participou e submeteu suas soluções, porém não foi considerado para os cálculos de média e desvio padrão em conjunto com os laboratórios acreditados por se tratar de outro grupo e não estarem inseridos no mesmo contexto de APS dos outros participantes. Desta forma, este trabalho apresenta na Tabela 2 apenas os resultados dos laboratórios acreditados, o que permite uma comparação com a primeira rodada.

Tabela 2 - Resultados dos Laboratórios Acreditados na Comparação da Segunda Rodada. [Inmetro, 2020]

ID	Escore de Mutação	Mutantes Mortos	Tamanho do bytecode (KB)	R_i
ID-2-c1d567723930371ff2060541e6d31885	2,11%	62	49,60	0,00043
ID-3-450ce5e8d70709a76f04c1016760bd49	0,00%	0	9,60	0,00000
ID-4-d0ab5067b86283a0d51db765dfd281e1	62,36%	1832	6,51	0,09513

Em função dos resultados apresentados na Tabela 2, tem-se que a média (μ) e o desvio padrão (σ) associados aos valores dos índices de desempenho (R_i) foram, respectivamente, 0,03185 e 0,05480. Portanto, a aplicação dos critérios estabelecidos, indica que todos os participantes apresentaram resultado satisfatório, exceto pelo

DOI: 10.17648/wrac-wciber-2020-125131
https://proceedings.science/p/125131?lang=en

participante identificado por “ID-4-d0ab5067b86283a0d51db765dfd281e1”, o qual teve seu desempenho classificado como “bom”.

É importante observar que o resultado 0 (zero) obtido por um dos participantes ocorreu porque o conjunto de casos de teste entregue apresentou erros durante a execução e, por consequência, não foi válido durante a execução da ferramenta PIT.

Outros pontos importantes foram evidenciados pelos participantes em seus Relatórios de Ensaio. Dentre esses pontos, destacam-se:

1. O tempo foi considerado curto para realizar a inspeção do código fonte e execução do desafio;
2. Também foi considerado que a tarefa de implementação de testes unitários e de mutação tem pouca relação com as atividades desempenhadas pelo laboratório, assim como, com avaliações no âmbito de Segurança Cibernética;
3. Foi evidenciado um problema na maneira como a ferramenta PIT identificava os mutantes mortos, onde, dependendo da abordagem adotada pelo participante para o desenvolvimento dos casos de teste, a ferramenta PIT podia marcar o mutante como vivo ou como morto por RUN_ERROR.

Em relação ao item 3, vale ressaltar que os resultados para as duas abordagens foram avaliados, porém, devido às regras adotadas no protocolo, não resultou em uma mudança da classificação do desempenho do participante. Os resultados considerando a segunda abordagem do participante, são apresentados na Tabela 3, onde os valores de média (μ) e desvio padrão (σ) foram, respectivamente, 0,01302 e 0,02218.

Tabela 3 - Resultados da comparação desconsiderando mutantes mortos marcados como RUN_ERROR. (Adaptado de [INMETRO, 2020])

ID	Escore de Mutação	Mutantes Mortos	Tamanho do bytecode (KB)	R_i
ID-2-c1d567723930371ff2060541e6d31885	2,11%	62	49,60	0,00043
ID-3-450ce5e8d70709a76f04c1016760bd49	0,00%	0	9,60	0,00000
ID-4-d0ab5067b86283a0d51db765dfd281e1	25,32%	744	6,51	0,03863

4. Análise Comparativa das Rodadas

Na primeira rodada, a questão relacionada à dificuldade dos testes, apontava na direção de que o desafio proposto na rodada do EP estava abaixo da competência técnica dos participantes. Tal conclusão é confirmada pelo feedback recebido de pelo menos 4 participantes em seus respectivos Relatórios de Ensaio. Eles alegam que os testes realizados foram de dificuldade baixa ou média. Por um lado, esse resultado é positivo por demonstrar que todos os participantes possuíam competência minimamente esperada na realização de tarefas imprescindíveis a uma APS. Por outro lado, o mesmo resultado indica que as próximas rodadas do EP deveriam propor cenários de teste mais desafiadores.

De forma complementar, pode-se assumir que o tempo destinado à realização do desafio proposto na primeira rodada foi suficiente para que os participantes pudessem exercitar diferentes casos de teste até que se identificasse o conjunto mais adequado que cobrisse todas as linhas do código-fonte identificadas nos relatório de referência,

DOI: 10.17648/wrac-wciber-2020-125131
https://proceedings.science/p/125131?lang=en

contribuindo mais uma vez para que todos os participantes obtivessem resultados semelhantes.

Na segunda rodada os resultados obtidos conseguiram discriminar o desempenho dos participantes, demonstrando que a metodologia baseada no escore de mutação contribuiu para melhorar a avaliação dos participantes. O grande número de classes a serem avaliadas e um curto espaço de tempo (4 dias) para desenvolvimento dos casos de teste também foram fatores importantes para garantir que os participantes não exercitassem ao máximo diferentes casos de teste, o que permitiria a todos obterem classificações máximas na rodada e levaria a um resultado semelhante ao encontrado na primeira rodada. Contudo, tendo em vista as regras adotadas para a avaliação do desempenho e uma baixa adesão de participantes, não foi possível obter uma classificação mais adequada dos participantes, o que pode ser observado na classificação do desempenho de participantes com escore de mutação igual a zero como "satisfatório".

Por fim, na seção de resultados da segunda rodada, foram relatadas inconsistências nos resultados apresentados pela ferramenta PIT, o que gerou dúvidas em um dos participantes durante o desenvolvimento dos casos de teste, apontando a necessidade de melhorias em relação às ferramentas utilizadas nos procedimentos estabelecidos para as comparações interlaboratoriais.

5. Considerações finais

A elaboração de regras e procedimentos para comparações interlaboratoriais no âmbito de APS é um trabalho em desenvolvimento que tem se mostrado desafiador, ao mesmo tempo que é uma iniciativa pioneira no Brasil, nesta área de conhecimento.

Os resultados da segunda rodada demonstraram que a metodologia de testes de mutantes contribuiu para elevar a complexidade dos desafios, permitindo classificar os desempenhos dos participantes em diferentes níveis e melhorando os resultados em comparação com a primeira rodada. No entanto, as regras estabelecidas no protocolo não permitiram atribuir uma classificação mais adequada para os participantes de acordo com seus respectivos resultados. Contudo, é importante ressaltar a competência demonstrada por todos os participantes em relação à linguagem Java, que aliada ao prazo, considerado curto pelos participantes, podem ter sido os principais motivos para a baixa participação na segunda rodada. Estes pontos e os feedbacks relatados pelos participantes demonstram as necessidades de melhoria a serem implementadas para uma próxima rodada de APS.

Como trabalhos futuros, será proposto (i) o desenvolvimento de desafios que permitam avaliar requisitos relacionados com a segurança cibernética dos itens de ensaio; (ii) o estabelecimento de requisitos para o desenvolvimento de um *software* que poderá ser utilizado como item de ensaio e esteja mais relacionado com as atividades desempenhadas pelos laboratórios acreditados; e por fim, (iii) a revisão das regras estabelecidas nos protocolos das comparações interlaboratoriais, de forma permitir uma classificação mais adequada dos desempenhos dos participantes.

6. Referências

ABNT NBR ISO/IEC 17043, Avaliação de conformidade - Requisitos gerais para ensaios de proficiência, ABNT, Rio de Janeiro, 2011.

DOI: 10.17648/wrac-wciber-2020-125131
https://proceedings.science/p/125131?lang=en

ANSSI (2015). Licensing of evaluation facilities for the first level security certification. 1.2 edition.

Câmara, S., Barras, T., Melo, W., Chapetta, W., e Machado, R. (2019). "Estabelecimento de um Pacote de Ensaio de Proficiência para avaliação de laboratórios em análise de produtos de software".

Delamaro, M.E, Maldonado, J.C e Jino, M. (2016). "Introdução ao teste de software", 2ª Edição, Elsevier.

INMETRO (2019). Relatório Final da Primeira Rodada de Ensaio de Proficiência. https://sicccIber.com.br/wp-content/uploads/2020/09/relatorio-final-ep-de-software-1-rodada-1.pdf [acesso: 10/12/2020]

INMETRO (2020). Relatório Final da Primeira Rodada Pública de Proficiência em Avaliação da Conformidade de Produto de Software. https://sicccIber.com.br/wp-content/uploads/2020/12/RelatorioFinal-1a-Rodada-Publica.pdf [acesso: 10/12/2020]

ISO 13528, Statistical methods for use in proficiency testing by interlaboratory comparison, ISO, Geneva, 2015.

Jaccard, P. (1912). "The distribution of the flora in the alpine zone". 1. New Phytologist, v. 11, n.2, p. 37-50.

Machado, R., Melo, W., Bento, L., Camara, S., da Hora, V., Barras, T., & Chapetta, W. (2020). Proficiency Testing for Software Analysis and Cybersecurity Laboratories. In *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT* (pp. 441-446). IEEE.

Maldonado, J. C. (1998). Aspectos teóricos e empíricos de teste de cobertura de software. ICMSC-USP.

Milano, D. T. (2011). "Android Application Testing Guide. Packt Publishing. ISO (2010). ISO/IEC 17043:2010 Conformity assessment – General requirements for proficiency testing. International Standard Organization, 1st edition.

Myers, G. J., Sandler, C., Badgett, T. (2011) "The art of software testing". John Wiley & Sons.

NIST (2017). National Voluntary Laboratory Accreditation Program (NVLAP). https://www.nist.gov/nvlap. [acesso: 27-06-2019].

DOI: 10.17648/wrac-wciber-2020-125131
https://proceedings.science/p/125131?lang=en