# Existem Concordância e Confiabilidade na Avaliação da Criatividade de Resultados Tangíveis da Aprendizagem de Computação na Educação Básica?

Nathalia da Cruz Alves
Departamento de Informática e Estatística
Universidade Federal de Santa Catarina
Florianópolis/Santa Catarina/Brasil
nathalia.alves@posgrad.ufsc.br

Christiane Gresse von Wangenheim
Departamento de Informática e Estatística
Universidade Federal de Santa Catarina
Florianópolis/Santa Catarina/Brasil
c.wangenheim@ufsc.br

Lúcia Helena Martins-Pacheco
Departamento de Informática e Estatística
Universidade Federal de Santa Catarina
Florianópolis/Santa Catarina/Brasil
lucia.pacheco@ufsc.br

Adriano Ferreti Borgatto
Departamento de Informática e Estatística
Universidade Federal de Santa Catarina
Florianópolis/Santa Catarina/Brasil
adriano.borgatto@ufsc.br

## RESUMO

À medida que uma sociedade cada vez mais globalizada cria economias baseadas no conhecimento, a necessidade de promover a criatividade nas escolas se intensifica. Embora a criatividade seja tipicamente abordada nas artes, literatura ou música, o ensino de computação também pode ser uma alternativa, especialmente ao estimular a criação de novos artefatos de software. Atualmente existe uma vasta pesquisa sobre criatividade, no entanto, a pesquisa para avaliar o grau de criatividade de produtos de software, criados como resultados tangíveis de aprendizagem, é pouco explorada. Ainda assim, como a pesquisa em geral sobre criatividade indica, uma questão central é: é possível fornecer avaliações sobre o grau de criatividade de aplicativos móveis desenvolvidos como resultados tangíveis de aprendizagem no ensino de computação na Educação Básica com concordância e confiabilidade? Assim, este artigo relata os resultados de uma análise de avaliações de 24 avaliadores sobre 10 aplicativos criados com App Inventor com relação à concordância e confiabilidade entre avaliadores. Além disso, também é discutido se tipos específicos de expertise impactam a percepção da criatividade. Resultados indicam ausência de concordância e confiabilidade entre avaliadores, e apontam para a importância do treinamento do avaliador em criatividade e/ou suporte automatizado de avaliação para obtenção de resultados mais confiáveis.

## PALAVRAS-CHAVE

Criatividade, Avaliação, Concordância, Confiabilidade.


*EduComp'21, Abril 27–30, 2021, Jataí, Goiás, Brasil (On-line)*


## 1 INTRODUÇÃO

O estímulo do pensamento criativo na Educação Básica e a capacidade de solucionar problemas de forma inovadora têm sido reconhecidos como componentes essenciais da cognição humana cujo fomento é fundamental para uma economia globalizada e para a educação de cidadãos globalmente competitivos [56][28]. Portanto, é essencial abordar a aprendizagem da criatividade como uma das habilidades do século XXI [18] para que os alunos possam solucionar problemas tecnológicos, do presente e futuro, de forma inovadora e criativa [18][43]. Apesar de a criatividade ser tipicamente ensinada como parte das artes, música e literatura, a educação em computação também oferece um ambiente potencialmente fértil para o desenvolvimento das habilidades de resolução de problemas e desenvolvimento criativo dos alunos [43]. A criatividade envolve um conjunto de processos de pensamento que se sobrepõem aos fundamentos da computação, como a observação, imaginação, visualização, abstração, criação e identificação de padrões, que podem apoiar o desenvolvimento da criatividade [59]. Especialmente, enfocando a aprendizagem baseada em problemas e atividades abertas com o objetivo de resolver problemas, o ensino de computação pode fornecer aos alunos um ambiente propício à promoção da criatividade [47]. Embora a necessidade de promover a criatividade na Educação Básica tenha sido estabelecida, o fomento da criatividade e das habilidades criativas de resolução de problemas podem ser desafiadoras em meio às expectativas de objetivos de aprendizagem explícitos e resultados mensuráveis dentro da sala de aula [12]. Com o foco crescente do ensino baseado em evidências, é fundamental haver maneiras válidas e confiáveis de avaliar a aprendizagem dos alunos [24]. Porém, mesmo existindo uma vasta área de pesquisa voltada à criatividade de forma geral [57] observa-se comumente uma falta de consenso sobre sua definição e maneiras de avaliação [30][52][57], especialmente quando o foco é

sobre artefatos de software criados como resultados de aprendizagem.

Trazendo a definição do artefato de software como produto criativo para a prática da sala de aula resulta na necessidade de que professores sejam capazes de compreender profundamente o que é criatividade e se ela está ou não presente nos artefatos dos alunos. No entanto, pesquisas que abordam a avaliação da criatividade por parte dos professores indicam que eles podem enfrentar desafios no reconhecimento da criatividade [11]. Nesse contexto, uma dimensão da avaliação da criatividade é a adequação das medições observacionais que dependem de avaliadores humanos. Essa adequação compreende a concordância e a confiabilidade entre avaliadores [22].

Atualmente, existem diversos estudos sobre a concordância e confiabilidade das avaliações da criatividade, no entanto, sua validação tem se limitado, principalmente, a avaliar a criatividade de artefatos produzidos sob condições experimentais rigidamente restritas com instruções uniformes fora do contexto de ensino de computação [8][9].

Neste contexto, este artigo apresenta resultados de um estudo comparando avaliações de 24 avaliadores de diferentes áreas pertinentes ao domínio sobre dez aplicativos que representam resultados tangíveis de aprendizagem no contexto do ensino de computação na Educação Básica. São analisadas a concordância e a confiabilidade entre os avaliadores. Considerando a importância da experiência do avaliador, é analisado também o impacto de variáveis intraindividuais relacionadas à expertise que podem influenciar a avaliação da criatividade.

# 2 FUNDAMENTAÇÃO TEÓRICA

## 2.1 Ensino de computação visando o desenvolvimento da criatividade

Uma das alternativas para desenvolver a criatividade é por meio do ensino de computação, por exemplo, por meio de desenvolvimento de aplicativos. Especificamente no contexto de aprendizagem baseada em problemas, a adoção de abordagens como a ação computacional [55] pode ser feita, por exemplo, ensinando os alunos a desenvolverem aplicativos móveis com App Inventor. Desta forma, são abordados não só conceitos de algoritmos e programação, mas também de design *thinking*, design de interface, etc. [19]. Como parte da progressão da aprendizagem, seguindo um ciclo como o *Use-Modifique-Crie* [36], especialmente durante o estágio de criação, os alunos são estimulados a desenvolver artefatos criativos de software. Esse tipo de progressão orienta os alunos na etapa *Use* a seguir instruções passo-a-passo para desenvolver um aplicativo. Na etapa *Modifique*, os alunos modificam o aplicativo, por exemplo, alterando/adicionando novas funcionalidades, design de interface do usuário, etc. Na etapa *Crie*, os alunos são desafiados, seguindo a abordagem baseada em problemas, a desenvolver seus próprios aplicativos com temas relevantes para a sua comunidade local. Desta forma, a partir de problemas do mundo real e de fatores que fazem parte do cotidiano do estudante o ensino é direcionado à busca de soluções, agregando-se ao conhecimento anterior do estudante (sua realidade cotidiana – conhecimento subsunçor), possibilitando abstração por meio de analogias e similaridades, favorecendo o desenvolvimento de estruturas cognitivas associadas à solução de problemas computacionais por meio do desenvolvimento de aplicativos. O aprendizado, dessa forma, passa a ser então significativo na concepção de Ausubel [41].

## 2.2 Avaliação da criatividade

A avaliação da criatividade e suas dificuldades têm sido debatidas há anos [58][26][48][6]. A dificuldade de medir e avaliar a criatividade está presente não só no ensino e aprendizagem da criatividade, mas nas diversas questões relacionadas aos constructos para representá-la na Psicologia, como apontado por Sternberg [52]. Dentro da literatura, existem diversas definições da criatividade que dependem do que está sendo avaliado [57]. Visando um melhor entendimento da criatividade, diversos pesquisadores têm proposto diversas formas de estruturar a definição de criatividade. Uma delas é a definição amplamente utilizada dos Quatro Ps de Rhodes [46] (Figura 1).


**Figura 1: Quatro Ps da criatividade [46]**

No contexto da educação, focando no ensino baseado em evidências, a discussão geralmente se concentra na criatividade do produto (Figura 1). Nesse contexto, diversos pesquisadores consideram que as melhores medidas de criatividade seriam por meio da análise de produto em cada nível de realização criativa [52]. Assim, a construção de instrumentos de avaliação baseada no desempenho da aprendizagem de criatividade dos alunos é com base nos artefatos criados como resultados tangíveis de aprendizagem. Focando especificamente na vertente do produto, a criatividade é tipicamente definida como sendo composta por três subfatores [9][54][14]: **originalidade** referindo-se a um produto original e novo [10], **utilidade** referindo-se a um produto apropriado que atende às necessidades práticas de um problema [42], e **condensação** referindo-se a um produto bem-feito que unifica os conceitos polares de simplicidade e complexidade [27].

A avaliação da criatividade do produto por professores requer uma profunda compreensão sobre o que é criatividade e se ela está ou não presente nos artefatos dos alunos. No entanto, pesquisas indicam que professores podem ter dificuldades para reconhecê-la [11]. Estudos têm mostrado que tais avaliações com base nos resultados de aprendizagem dos alunos muitas vezes não são confiáveis, exigindo, até mesmo para modelos maduros de avaliação da criatividade, avaliadores experientes no domínio específico, a fim de avaliar os produtos de uma forma confiável e válida [11][30]. Por outro lado, em muitos estudos a avaliação é feita de forma subjetiva, sem um modelo de avaliação como suporte [4][32][51]. Alguns desses estudos indicam que a avaliação

subjetiva parece ser capaz de produzir avaliações consistentes [30]. Essas avaliações subjetivas são baseadas na definição de Amabile [4] que argumenta que algo é criativo na medida em que as pessoas concordam que é. Ao adotar tal abordagem subjetiva, pode-se considerar que a criatividade do produto é tipicamente avaliada com base em seus principais subfatores — originalidade, utilidade e condensação conforme definidas anteriormente, de modo que a restrição e a definição de subfatores auxiliam na delimitação do foco de avaliação.

## 2.3 Adequação da avaliação da criatividade

A adequação da avaliação da criatividade, quando realizada por meio de medições observacionais que dependem de avaliadores humanos, inclui a concordância e a confiabilidade entre avaliadores [22].

**Concordância** entre avaliadores refere-se ao grau em que dois ou mais avaliadores usando a mesma escala de classificação dão a mesma classificação a um produto observável idêntico [22]. Assim, a concordância entre avaliadores é uma medida de consenso entre as classificações atribuídas pelos avaliadores.

**Confiabilidade** entre avaliadores, por outro lado, é definida como a medição da consistência entre os avaliadores [22]. Essa consistência se refere à ordenação ou posição relativa das classificações atribuídas por cada avaliador. Assim, se as avaliações de criatividade de um avaliador são sempre mais altas do que as avaliações de outros avaliadores, ele é consistente, mas não concordante com os demais (Figura 2).

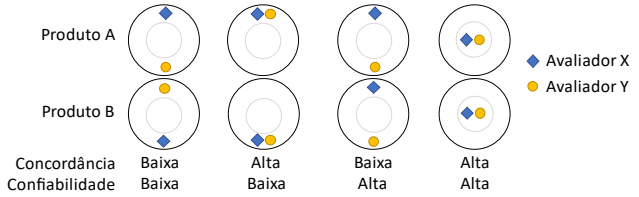


**Figura 2: Concordância e confiabilidade de avaliações supondo que o Produto A é idêntico ao produto B**

Alta concordância e confiabilidade entre avaliadores são desejáveis para dar credibilidade aos resultados da avaliação. Se a concordância for baixa, isso indica que as pontuações dependem muito dos indivíduos que realizaram as avaliações. Se a confiabilidade for baixa, isso indica que diferentes critérios subjetivos estão sendo usados e, portanto, alguns alunos podem estar sendo prejudicados na avaliação. Assim, a concordância e a confiabilidade entre avaliadores são essenciais para garantir que as avaliações sejam precisas e significativas [25][49]. Sem demonstrar que dois avaliadores independentes podem ser treinados para avaliar de forma consistente e concordante um determinado produto, a possibilidade de alcançar a medição objetiva dos fenômenos educacionais é diminuída [34].

## 2.4 Estudos existentes sobre a adequação da avaliação da criatividade

Existem diversos estudos sobre a adequação da avaliação da criatividade, no entanto, sua validação tem se limitado principalmente a avaliar a criatividade de artefatos produzidos sob condições experimentais rigidamente restritas com instruções uniformes (Tabela 1). Embora tenham sido obtidos níveis altos de confiabilidade entre avaliadores nesses experimentos [5] não está claro se esses resultados também seriam válidos no contexto da educação em computação cujos resultados de aprendizagem "reais" na forma de artefatos computacionais complexos são produzidos sob condições menos controladas. Tipicamente, artefatos computacionais criados por alunos em resposta a instruções amplamente variadas são realizados dentro de contextos de aprendizagem construtiva sem quaisquer soluções e/ou processos predefinidos.

**Tabela 1: Estudos sobre a avaliação da criatividade**

| Nome e referência | Domínio | Avaliadores | Adequação da avaliação | |
|---|---|---|---|---|
| | | | Concordânc. | Confiabili. |
| *Student Product Assessment Form* [45] | Produtos científicos, estudos sociais, audiovisuais, e textos criativos | 4 professores | ~0.80 a 95.5 | 0.39 a 1 |
| *Creativity Product Inventory* [54] | Produtos construídos com materiais selecionados | 2 avaliadores treinados | - | 0.87 a 0.97 |
| *Creative Product Semantic Scale* [8][9] | Camisas | 133 alunos | - | 0.69 s 0.91 |
| | Cadeiras | 128 alunos | - | 0.77 a 0.87 |
| *Invention Evaluation Scale* (IES) [58] | Invenções livres usando sucata | Alunos e recém-formados de programas de doutorado | 0.69 a 0.92 | 0.94 a 0.91 e 0.50 a 0.63 |
| *Creative Solution Diagnosis Scale* (CSDS) [14] | Modelos de veículos com rodas | 13 professores | - | 0.79 |

Considerando que a avaliação da criatividade é dependente de domínio [29], há uma falta de pesquisas enfocando a concordância e confiabilidade da avaliação da criatividade de artefatos de software. Pesquisas existentes que estudam a concordância e a confiabilidade entre avaliadores no contexto de computação, como El-Emam et al. [17] ou Gresse von Wangenheim et al. [23], não enfocam a criatividade, mas outros temas de avaliação, como a estética visual de aplicativos [23]. Por outro lado, pesquisas que enfocam a criatividade no contexto da computação [50][39][1] não abordam a questão da concordância e confiabilidade da avaliação. Assim, não foram encontradas pesquisas sobre a concordância e confiabilidade da avaliação da criatividade de artefatos de software, como aplicativos móveis, desenvolvidos como resultados de aprendizagem no contexto da educação em computação na Educação Básica.

Estudos existentes sobre avaliação da criatividade em diversos domínios argumentam que avaliadores que fornecem avaliações subjetivas de criatividade devem ter alguma experiência no domínio. No contexto do ensino de computação na Educação Básica, tal requisito pode ser difícil de alcançar, pois, atualmente,

há uma grande falta de professores de computação neste estágio educacional, e o ensino de computação é muitas vezes introduzido de forma interdisciplinar por professores de outras áreas [2][3].

Na avaliação educacional é esperado que avaliadores concordem em suas avaliações. Se diferentes avaliadores chegam a diferentes conclusões sobre a criatividade de um produto que estão avaliando, não se pode saber em quais avaliações confiar. Assim, deve-se garantir que as avaliações sejam resultados de um processo sistemático e não atribuídas aleatoriamente. Para que essas avaliações sejam confiáveis é necessário que haja concordância e confiabilidade entre os avaliadores.

# 3 METODOLOGIA DE PESQUISA

## 3.1 Definição da pesquisa

Com o objetivo de analisar a adequação da avaliação da criatividade do produto focando em aplicativos móveis, são analisadas a concordância e a confiabilidade entre avaliadores. Assim, a pergunta de pesquisa deste estudo é se existem concordância e confiabilidade na avaliação da criatividade de resultados tangíveis da aprendizagem de computação na Educação Básica. Nesse contexto, a concordância é o grau em que dois ou mais avaliadores usando a mesma escala de classificação atribuem uma mesma avaliação para um aplicativo e a confiabilidade é definida como a medição da consistência entre avaliadores [37]. Assim, são analisadas as seguintes questões de pesquisa:

**QP1**. A criatividade do produto focando em aplicativos pode ser avaliada com concordância e confiabilidade entre avaliadores?

**QP2**. Quais variáveis intraindividuais dos avaliadores podem influenciar na concordância e/ou confiabilidade da avaliação entre avaliadores?

A análise foca na quantificação da concordância e confiabilidade entre as avaliações ordinais de avaliadores sobre a criatividade do produto e seus subfatores. Essa quantificação permite medir o grau de concordância e confiabilidade das avaliações dadas pelos avaliadores.

Para analisar a concordância e confiabilidade entre avaliadores, é conduzido um estudo exploratório adotando um design totalmente cruzado (*fully crossed design*) [26] no qual todos os resultados de aprendizagem são avaliados por todos os avaliadores. Assim, é avaliada a criatividade do produto e seus subfatores (originalidade, utilidade e condensação) de aplicativos móveis criados com App Inventor no contexto do ensino de computação na Educação Básica (Figura 3).

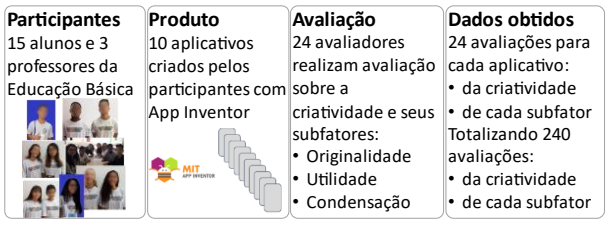


**Figura 3: Visão geral do estudo exploratório**

Os aplicativos são resultados de experiências da Iniciativa Computação na Escola nas quais alunos e professores da Educação Básica desenvolveram aplicativos como parte do ensino de computação. Como resultado, 10 aplicativos foram selecionados com vários graus de criatividade do produto. Os aplicativos foram apresentados aos avaliadores com a descrição da motivação para o desenvolvimento do aplicativo e seus criadores, bem como os objetivos, funcionalidades e capturas de tela dos aplicativos (Figura 4). Além disso, links para o código-fonte do App Inventor e arquivo APK foram fornecidos para uma análise detalhada possibilitando a experimentação do produto.

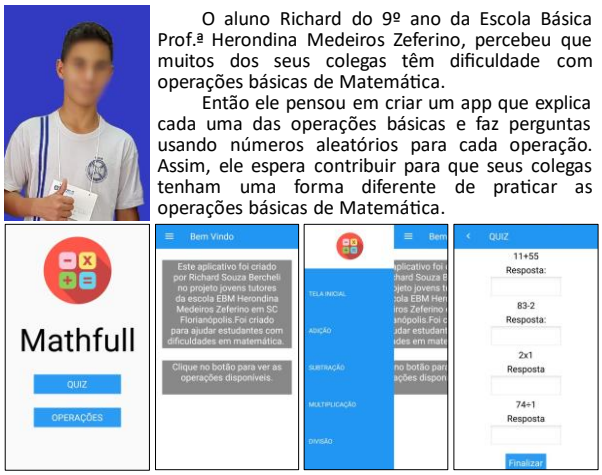


**Figura 4: Extrato de um exemplo de apresentação de um aplicativo (MathFull) como parte das instruções**

Todos os participantes do estudo receberam as mesmas instruções, incluindo uma breve visão geral sobre a definição de criatividade do produto e seus subfatores: originalidade, utilidade e condensação. Todos as avaliações foram feitas com uma escala ordinal de 3 pontos (pouco criativo; mais ou menos criativo e muito criativo). Foi solicitado aos avaliadores que após a avaliação de cada aplicativo, indicassem quais características mais influenciaram a avaliação.

Com o objetivo de analisar a influência de variáveis intraindividuais sobre a concordância e confiabilidade entre avaliadores, foram coletadas informações sobre as áreas de expertise e tempo de experiência com o domínio (Tabela 2).

**Tabela 2: Variáveis intraindividuais analisadas**

| Variáveis intraindividuais | Possíveis respostas |
|---|---|
| Área(s) de expertise (seleção múltipla possível) | Ciência da Computação; Design; Educação em computação; Pedagogia; Psicologia |
| Experiência de ensino/pesquisa de computação na Educação Básica (anos) | Nenhum; 1 a 2 anos; 3 a 5 anos; Mais do que 5 anos |
| Experiência de pesquisa sobre a criatividade (anos) | Nenhum; 1 a 2 anos; 3 a 5 anos; Mais do que 5 anos |
| Experiência com desenvolvimento de apps com App Inventor (quantidade) | Nenhum; 1 a 2 apps; 3 a 5 apps; Mais do que 5 apps |

## 3.2 Execução

Considerando as restrições de mobilidade devido a pandemia da COVID-19, o estudo foi realizado de forma online. Os participantes receberam uma explicação sobre o estudo e, em seguida, foram dadas as instruções para a tarefa de avaliação.

Foram convidados um total de 29 avaliadores com formação em diferentes áreas. Dentre esses, 24 responderam à pesquisa representando uma taxa de resposta de 82%. Os participantes incluíram pesquisadores da área de educação em computação, professores do Ensino Superior e professores da Educação Básica, bem como alunos de graduação e pós-graduação.

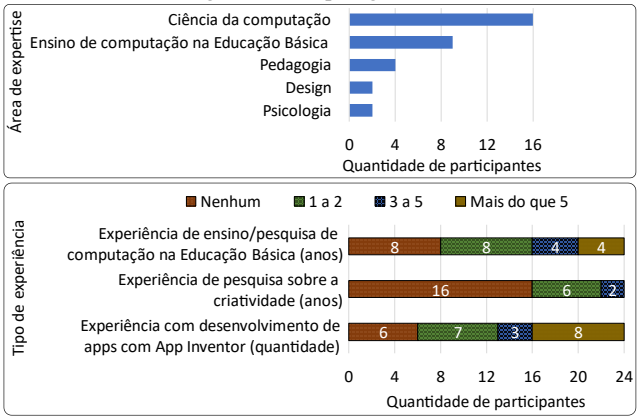


**Figura 5: Expertise e experiência dos participantes**

Em geral os participantes possuem alguma experiência de ensino de computação na Educação Básica ou com o desenvolvimento de aplicativos com App Inventor (Figura 5). No entanto, poucos participantes têm experiência em pesquisa relacionada à criatividade.

Os participantes analisaram e avaliaram independentemente a criatividade do produto de todos os 10 aplicativos usando uma escala ordinal de 3 pontos.

## 3.3 Análise dos dados

As medidas de concordância para escalas ordinais se concentram em quantificar os níveis de concordância entre os avaliadores, ou seja, analisar se cada avaliador atribui uma categoria idêntica a um resultado de aprendizagem. As análises de concordância entre avaliadores incluem várias versões do Kappa de Cohen, incluindo o Kappa de Fleiss [20] para avaliar a concordância entre vários avaliadores quando as avaliações usando escalas ordinais estão sendo examinadas, como no presente estudo. Tipicamente, valores de Kappa de Fleiss são inaceitáveis se kappa ≤ 0,40. Além disso, um valor de Kappa negativo indica forte discordância entre avaliadores e é considerado um sinal de baixa concordância (Figura 6) [22].

A confiabilidade pode ser medida via alfa de Cronbach, o qual foi originalmente desenvolvido como uma medida de confiabilidade em testes psicométricos, mas também é a medida amplamente usada de consistência interna de escalas e, recentemente, de confiabilidade entre avaliadores na literatura [33]. No contexto deste trabalho, a confiabilidade entre os avaliadores é analisada para identificar se existe um padrão das avaliações, mesmo tendo avaliadores mais rigorosos e menos rigorosos, ou seja, a avaliação não sendo igual. O objetivo é identificar a coerência no padrão das avaliações. Um valor alto de alfa de Cronbach significa que as avaliações são coerentes e têm confiabilidade excelente [13].

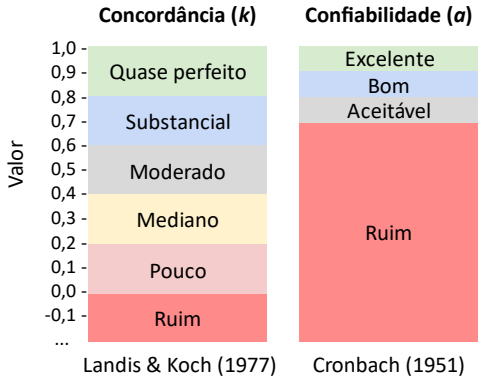


**Figura 6: Valores limiares para concordância com Kappa e confiabilidade com alfa de Cronbach**

A concordância e a confiabilidade entre avaliadores podem ser avaliadas tanto sobre a avaliação geral da criatividade, bem como sobre seus subfatores. A análise da concordância e da confiabilidade com relação aos subfatores fornece informações sobre quais subfatores os avaliadores podem ter dificuldades na avaliação [22]. Além disso, para analisar se determinadas áreas e/ou experiências influenciam a concordância e/ou confiabilidade entre avaliadores são calculados o Kappa de Fleiss [20] e o alfa de Cronbach [13] respectivamente para cada variável intraindividual, de forma a analisar sua influência nos resultados da avaliação.

# 4 RESULTADOS

## 4.1 A criatividade do produto pode ser avaliada com concordância e confiabilidade entre avaliadores?

Para analisar esta questão, foi analisada a concordância e confiabilidade entre avaliadores sobre a avaliação como um todo, reunindo todos os dados de avaliação da criatividade e seus subfatores. Além disso, também foi analisada a concordância e confiabilidade da criatividade do produto separadamente, bem como de cada um dos subfatores (originalidade, utilidade e condensação) separadamente (Tabela 3).

**Tabela 3: Concordância e confiabilidade da avaliação da criatividade do produto como um todo**

| Aspecto | Medida | Resultado (24 avaliadores) |
|---|---|---|
| Concordância | Kappa de Fleiss | 0,117 |
| Confiabilidade | Alfa de Cronbach | 0,89 |

Os resultados indicam que foi obtida boa confiabilidade por meio do alfa de Cronbach (a = 0,89). No entanto, a concordância entre os avaliadores obtida por meio do Kappa de Fleiss (0,117) demonstra que as avaliações não possuem concordância e variaram

amplamente. Essa falta de concordância pode ser visualizada na Figura 7 que apresenta a distribuição das avaliações na escala ordinal de 3 pontos em relação a avaliação dos 10 aplicativos pelos 24 participantes do estudo.

*4.1.1 Concordância e confiabilidade entre avaliadores com relação à criatividade do produto*

Analisando somente a avaliação da criatividade do produto desconsiderando seus subfatores (originalidade, utilidade e condensação), foi obtido um valor de Kappa de Fleiss muito baixo, indicando que não pode ser observada nenhuma concordância entre os participantes quanto à avaliação da criatividade do produto separadamente (Tabela 4).

**Tabela 4: Concordância e confiabilidade da avaliação da criatividade do produto desconsiderando seus subfatores**

| Aspecto | Medida | Resultado (24 avaliadores) |
|---|---|---|
| Concordância | Kappa de Fleiss | 0,115 |
| Confiabilidade | Alfa de Cronbach | 0,88 |

Porém, com base no alfa de Cronbach a confiabilidade pode ser considerada boa, indicando que os participantes avaliaram a criatividade dos aplicativos de forma consistente (Tabela 4).

De forma a identificar as razões para a falta de concordância entre os avaliadores sobre a criatividade do produto, foram analisadas as explicações dos avaliadores sobre quais características os levaram a chegar às suas avaliações de criatividade do produto (Tabela 5). Por meio dessa análise, observou-se que os avaliadores utilizaram diferentes critérios implícitos para a tomada de decisão sobre a avaliação, haja vista que neste estudo não foram predefinidas medidas detalhadas. A Tabela 5 apresenta um exemplo com uma visão geral sobre esses diferentes critérios com base nos comentários feitos pelos avaliadores explicando sua avaliação do aplicativo MathFull – um aplicativo de quiz sobre as 4 operações básicas de matemática.

**Tabela 5: Exemplos de explicações sobre a avaliação da criatividade do produto (aplicativo MathFull)**

| Tipo de avaliação | Descrição | Exemplos de avaliação do app MathFull (aplicativo de *quiz* sobre as 4 operações básicas de matemática) |
|---|---|---|
| Contextual | Baseada no contexto escolar/ pedagógico | "Ele é extremamente útil no contexto da escola." (Avaliador 21)<br>"Considero também que o tempo que o estudante tem disponível para desenvolver o aplicativo é um fator importante (...) Partindo da ideia de que é um primeiro protótipo, com o objetivo pedagógico de aprender o processo, avaliaria como excelente o resultado." (Avaliador 10) |
| | Baseada em características do aluno | "Na minha percepção o aplicativo é de fato criativo levando em conta a faixa etária do estudante (...)" (Avaliador 10)<br>"Apesar de possuir alguns deslizes em seu design, é bem apresentável em relação aos apps desenvolvidos por muitos estudantes neste ano da educação básica." (Avaliador 12) |
| Focada | Baseada somente em um fator (p. ex. originalidade) | "O app é pouco criativo porque existem outros com funcionalidades bem parecidas." (Avaliador 16)<br>"Não é uma ideia nova. Seria mais criativo se não fosse completamente aleatório, mas avaliasse o nível do usuário e sugerisse exercícios com grau crescente de dificuldade." (Avaliador 8) |
| Ponderada | Baseada em 2 fatores: utilidade e condensação | "Porque ele pode ajudar no entendimento sobre a matéria de matemática de forma diferente e intuitiva. Além disso, ele apresenta um design visual bem agradável." (Avaliador 4).<br>"Achei o app mais ou criativo pois ele é muito útil, mas achei que a combinação de cores não é atrativa." (Avaliador 11) |
| | Baseada em 2 fatores: originalidade e utilidade | "Considerei o app mais ou menos criativo, pois é possível encontrar alguns app com essa finalidade, no entanto, é bem didático e útil aos estudantes." (Avaliador 5) |
| | Baseada em 2 fatores: originalidade e utilidade | "Considerei original pois foi a primeira vez que vi um aplicativo com este objetivo. Considerei mais ou menos útil e condensado pois ele pode ser relevante para os alunos com dificuldade e existem coisas que podem melhorar no seu acabamento." (Avaliador 14)<br>"Não é original, mas tem suas especificidades, é útil pois tem um objetivo claro e é bem feito, considerando o contexto." (Avaliador 19) |
| Simplista | Baseada em comparação fora do domínio | "Não oferece muito mais do que uma calculadora ofereceria." (Avaliador 6) |

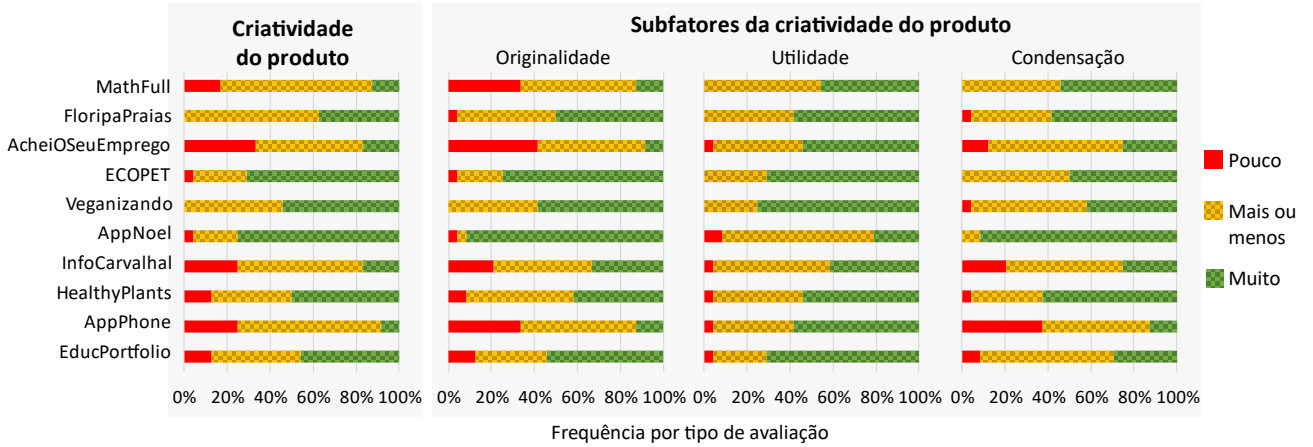


**Figura 7: Distribuição das avaliações da criatividade e seus subfatores (originalidade, utilidade e condensação)**

Alguns participantes demonstraram uma avaliação altamente focada ao analisar o aplicativo apenas em um dos subfatores da criatividade (principalmente originalidade). Isso pode ser explicado pelo fato de que tipicamente a originalidade é a principal característica para a percepção da criatividade, antecedendo a percepção de outros subfatores (utilidade e condensação) [27][60] que tipicamente são considerados secundários [16]. Outros avaliadores demonstraram uma avaliação mais holística, considerando implicitamente todos os três subfatores da criatividade do produto ou levando em consideração o contexto educacional e/ou características dos alunos, como a idade (Tabela 5).

Observou-se também que uma das explicações pode indicar um entendimento simplificado ou mesmo errôneo do aplicativo, comparando-o com calculadoras, sendo que o aplicativo MathFull é um aplicativo de quiz de matemática com finalidade educacional, e não com a finalidade de fazer cálculos. Isso também sugere que ao exigir a avaliação de um grande número de produtos/artefatos, problemas relacionados à fadiga e falta de atenção, podem impactar negativamente a concordância e/ou confiabilidade entre avaliadores, e, portanto, a qualidade da avaliação.

#### *4.1.2 Concordância e confiabilidade entre avaliadores com relação à originalidade, utilidade e condensação*

Com base na decomposição da criatividade nos subfatores de originalidade, utilidade e condensação, foram analisadas as avaliações feitas em cada um desses subfatores (Tabela 6). A Figura 7 também apresenta a distribuição das avaliações por cada subfator na escala ordinal de 3 pontos com relação aos 10 aplicativos avaliados pelos 24 participantes do estudo.

**Tabela 6: Concordância e confiabilidade das avaliações de originalidade, utilidade e condensação**

| **Subfator** | **Concordância** | **Confiabilidade** |
|---|---|---|
| | Kappa de Fleiss | Alfa de Cronbach |
| Originalidade | 0,147 | 0,92 |
| Utilidade | 0,043 | 0,71 |
| Condensação | 0,110 | 0,91 |

É possível observar que a concordância entre avaliadores com relação a todos os três subfatores permanece muito baixa (abaixo de 0,2), indicando que também não há um consenso de avaliação em qualquer um desses subfatores. Comparando-os com o valor do Kappa de Fleiss no que diz respeito à avaliação da criatividade do produto (seção 4.1.1), a concordância entre avaliadores com relação à originalidade dos aplicativos é um pouco maior, enquanto o valor de condensação é um pouco menor, e o referente à utilidade é quase zero, indicando que não há concordância (Tabela 6).

Por outro lado, analisando a confiabilidade entre avaliadores, os valores do alfa de Cronbach da avaliação da originalidade e condensação indicam excelente confiabilidade (Tabela 6), indicando que os participantes mantiveram uma avaliação consistente de originalidade e condensação entre os aplicativos. Porém, a confiabilidade com relação à utilidade apresentou apenas confiabilidade aceitável, com um valor muito próximo do mínimo 0,7 [13]. Esse resultado pode ser devido às suposições na literatura de que as dimensões de originalidade e utilidade podem ser ortogonais [38] e, portanto, precisam ser investigadas separadamente, pois os fatores que influenciam a originalidade muitas vezes não são os fatores que influenciam a utilidade.

Observando o alto valor do alfa de Cronbach na avaliação da originalidade (Tabela 6), o qual é similar ao valor da criatividade em geral, uma das razões pode ser devido à percepção da originalidade ser considerada primária para a percepção da criatividade [60]. Neste contexto, a avaliação é realizada de forma mais direta que os demais subfatores, por exemplo, comparando a ideia do aplicativo, funcionalidades, etc., com aplicativos existentes considerando um universo de produtos criados por pessoas com experiência e formação semelhantes [27]. Portanto, a avaliação da originalidade do produto pode ser menos complexa e mais fácil de manter a consistência resultando em alta confiabilidade.

Por outro lado, o baixo valor de concordância em geral, até mesmo na avaliação da originalidade (Tabela 6), pode ser porque mesmo entre produtos criativos, existem muitos graus de criatividade [31], como aqueles identificados no contexto do ciclo de *Use-Modifique-Crie* [36] ou no Modelo de Propulsão [53]. Os produtos criativos podem variar entre contribuições menores modificando e/ou remixando aplicativos na etapa *Modique*, a aplicativos completamente novos como parte da etapa *Crie*. Assim, apesar da alta confiabilidade na avaliação da originalidade, os resultados indicam baixa concordância em geral em relação à avaliação sobre o quão criativo/original/útil/condensado um aplicativo é.

Observando ainda a baixa concordância em geral (Tabela 6), a avaliação da criatividade e dos seus subfatores deve se basear em um universo de produtos criados por pessoas com experiência e formação semelhantes. No contexto educacional, focando no ensino do desenvolvimento de aplicativos com o App Inventor, isso pode ser representado, por exemplo, por aplicativos compartilhados na galeria do App Inventor, uma plataforma na qual aplicativos criados pela comunidade são compartilhados. No entanto, no contexto deste estudo, ainda permanece a questão sobre qual universo de comparação os avaliadores se basearam de fato na hora de fazer suas avaliações. Devido à quantidade de 10 aplicativos (e não apenas 1), é difícil identificar se, na hora de avaliar, em vez de basear-se em aplicativos criados por alunos dentro de um contexto educacional, os avaliadores utilizaram seus conhecimentos e gostos adquiridos ao longo de anos de uso de aplicativos semelhantes como parte de sua vida pessoal ou experiência de ensino.

Com relação a uma maior confiabilidade na avaliação da condensação do que da utilidade (Tabela 6), uma razão pode ser devido ao fato de que a condensação inclui aspectos como simplicidade e estética do aplicativo, incluindo design visual, funcionalidades, etc., que por se tratarem de aspectos visuais, podem ser mais simples de avaliar que a utilidade. Tipicamente, a utilidade é altamente influenciada pelo contexto e muitas vezes leva tempo, energia mental considerável e experimentação [60]. Assim, essa natureza dinâmica sugere que a percepção da utilidade pode ser mais complexa que a originalidade e condensação, e, consequentemente, pode-se esperar menor concordância e confiabilidade quanto à utilidade.

## 4.2 Quais variáveis intraindividuais podem influenciar na adequação da avaliação?

Considerando que a avaliação de criatividade (e seus subfatores) no contexto deste estudo envolve várias competências, foi realizada uma análise separada por cada competência. Para cada categoria de competência, foi calculado o Kappa de Fleiss para avaliar a concordância e o Alfa de Cronbach para avaliar a confiabilidade. Desta forma, é possível compará-las com a concordância geral (k=0,117) e confiabilidade em geral (a=0,89) apresentadas na Seção 4.1.

### *4.2.1 Área(s) de expertise*

De acordo com as respostas dos participantes, foram analisadas separadamente a concordância e confiabilidade para cada uma das cinco áreas de expertises indicadas pelos próprios participantes. Como os participantes puderam indicar mais do que uma área de expertise, a soma total dos avaliadores está acima do total de participantes (Tabela 7).

**Tabela 7: Concordância e confiabilidade da avaliação por área(s) de expertise**

| Área(s) de expertise | | Ciência da Computação | Design | Educação em computação | Pedagogia | Psicologia |
|---|---|---|---|---|---|---|
| **Avaliadores (24 no total)** | | 16 | 2 | 9 | 4 | 2 |
| Concordância | Kappa de Fleiss | 0,122 | 0,117 | 0,113 | -0,02 | -0,14 |
| Confiabilidade | Alfa de Cronbach | 0,85 | 0,52 | 0,77 | 0,17 | -0,10 |

Observa-se somente um valor de Kappa de Fleiss minimamente mais alto nas áreas de Design, Ciência da Computação e Educação em Computação do que em Pedagogia e Psicologia, porém ainda todos abaixo de 0,2 indicando baixa concordância em geral. As áreas de Pedagogia e Psicologia inclusive apresentam valores negativos de concordância e confiabilidade, indicando que expertise nestas áreas pode causar um maior grau de dissenso nas avaliações. Uma razão pode ser talvez pela visão mais multidimensional que os avaliadores dessas áreas podem ter sobre criatividade. No entanto, cabe ressaltar que, no presente estudo, essas áreas de expertise estão representadas por uma amostra de avaliadores muito pequena, e, portanto, futuros estudos com um número maior de representantes dessas áreas devem verificar esta indicação.

### *4.2.2 Experiência referente a ensino/pesquisa de computação na Educação Básica*

De acordo com as respostas dos participantes, foram analisadas separadamente a concordância e a confiabilidade para cada um dos intervalos de tempo de experiência na área de ensino/pesquisa de computação na Educação Básica (Tabela 8).

**Tabela 8: Concordância e confiabilidade da avaliação por experiência de ensino/pesquisa de computação na Educação Básica**

| Tempo de experiência de ensino/pesquisa de computação na Educação Básica | | Nenhum | 1 a 2 anos | 3 a 5 anos | Mais do que 5 anos |
|---|---|---|---|---|---|
| **Avaliadores** | | 8 | 8 | 4 | 4 |
| Concordância | Kappa de Fleiss | 0,11 | 0,123 | 0,181 | 0.09 |
| Confiabilidade | Alfa de Cronbach | 0,72 | 0,74 | 0,56 | 0,63 |

Observa-se que em geral o tempo de ensino/pesquisa de computação na Educação Básica não aumentou a concordância no contexto deste estudo. Inclusive, avaliadores que indicaram mais do que cinco anos de experiência nessa área apresentaram uma concordância menor do que avaliadores com um a cinco anos e até mesmo do que avaliadores com absolutamente nenhum tempo de experiência. Além disso, nota-se que avaliadores com nenhuma experiência ou um a dois anos apresentaram uma confiabilidade maior que os demais. Isso pode indicar que tempo de ensino/pesquisa de computação na Educação Básica não necessariamente beneficia uma maior concordância ou confiabilidade. Esse resultado está de acordo com algumas pesquisas de outras áreas que indicam que professores menos experientes parecem ter uma visão mais balanceada da criatividade do que professores com mais experiência [40]. No entanto, como em geral todos os intervalos de tempo de experiência na área de ensino/pesquisa de computação na Educação Básica apresentam baixa concordância, as concepções de criatividade dos professores podem estar incompletas em geral, conforme indicado por Mullet et al. [40]. Neste contexto, uma alternativa pode ser treinar professores a fim de melhorar sua compreensão acerca da criatividade dentro do ensino de computação, bem como a compreensão sobre criatividade no contexto educacional como um todo [40].

### *4.2.3 Experiência com desenvolvimento de apps com App Inventor*

Considerando que a avaliação foi feita sobre produtos focando em aplicativos criados com App Inventor e que conhecimento do domínio é relevante, foram analisadas separadamente a concordância e a confiabilidade para diferentes quantidades de apps que cada participante já desenvolveu com App Inventor.

**Tabela 9: Concordância e confiabilidade da avaliação por experiência com desenvolvimento de apps com App Inventor**

| Quantidade de apps desenvolvidos com App Inventor | | Nenhum | 1 a 2 apps | 3 a 5 apps | Mais do que 5 apps |
|---|---|---|---|---|---|
| **Avaliadores** | | 6 | 7 | 3 | 8 |
| Concordância | Kappa de Fleiss | 0,052 | 0,143 | 0,036 | 0.123 |
| Confiabilidade | Alfa de Cronbach | 0,59 | 0,73 | 0,38 | 0,77 |

Como nas análises anteriores, nota-se que a quantidade de apps desenvolvidos com App Inventor não parece aumentar sistematicamente a concordância, pois todos apresentam um valor de Kappa de Fleiss abaixo de 0,2 (Tabela 9). A confiabilidade também não parece apresentar um aumento de acordo com a quantidade de apps desenvolvidos no contexto deste estudo. Observa-se também uma menor confiabilidade nos avaliadores que criaram de três a cinco apps, contudo este intervalo possui o menor número de avaliadores. Isso pode ter influenciado o resultado do alfa de Cronbach, pois à medida que se aumenta o número de indivíduos (neste caso, avaliadores), maior é a variância esperada, de tal forma que se obtém um valor superestimado da confiabilidade da avaliação quanto maior o número de avaliadores [35]. Portanto, uma confiabilidade maior nos demais intervalos de quantidade de apps pode estar superestimada por ter um número maior de participantes do que no intervalo de três a cinco apps.

*4.2.4 Experiência referente à criatividade*

De acordo com as respostas também foram analisadas separadamente a concordância e confiabilidade para cada um dos quatro intervalos de tempo de experiência referente à criatividade (Tabela 10). De forma geral, mesmo reconhecendo a criatividade como uma das habilidades importantes no século XXI observa-se atualmente ainda uma deficiência em geral sobre esta habilidade e principalmente como avaliá-la, como este assunto atualmente é raramente abordado na formação de professores. Essa falta de experiência referente à criatividade também se reflete na população do presente estudo em que a maioria não tem expertise especificamente relacionada a área de criatividade e o máximo de expertise observado foi entre 3-5 anos.

**Tabela 10: Concordância e confiabilidade da avaliação por tempo de experiência de pesquisa sobre a criatividade**

| Experiência de pesquisa sobre a criatividade | | Nenhum | 1 a 2 anos | 3 a 5 anos | Mais do que 5 anos |
|---|---|---|---|---|---|
| | **Avaliadores** | 16 | 6 | 2 | 0 |
| Concordância | Kappa de Fleiss | 0,089 | 0,183 | 0,205 | -- |
| Confiabilidade | Alfa de Cronbach | 0,79 | 0,77 | 0,36 | -- |

Observa-se que, conforme o tempo de experiência aumenta, a concordância entre os avaliadores também aumenta (Tabela 10). No entanto, o aumento no Kappa de Fleiss' referente ao maior tempo de experiência com criatividade observada na amostra (três a cinco anos) pode ser causada pelo fato de que a análise do presente estudo foi realizada com uma amostra de somente dois avaliadores. Esse número pequeno também pode ter influenciado o cálculo do alfa de Cronbach que tende a ser menor quando há menos participantes [35]. Um maior tempo de experiência em criatividade apresentou uma pequena melhora na concordância e confiabilidade, porém como a amostra contém poucos participantes com tempo de experiência considerável nessa área, esta questão em aberto precisa ser verificada em estudos futuros.

Em geral os resultados indicam que um maior treinamento/formação e tempo de experiência especificamente na área da criatividade podem ser benéficos, resultando em avaliações mais coerentes e consistentes. No entanto, ainda que a abordagem desse tema na formação do professor seja essencial, também é necessário que se tenha ferramentas e modelos validados voltados à avaliação da criatividade, de forma a auxiliar o professor nesse processo. Considerando também a alta carga de trabalho do professor em geral, a adoção de abordagens de avaliação automatizada pode facilitar o trabalho do professor, apoiando a avaliação da criatividade.

# 5 DISCUSSÃO

Em geral, os resultados deste estudo indicam níveis muito baixos de concordância entre os avaliadores. Desta forma, a avaliação da criatividade do produto focando em aplicativos móveis como resultados de aprendizagem pode diferir muito entre avaliadores por sua subjetividade e dificuldade de medir a criatividade de tais artefatos de software de forma consistente.

Ao comparar os resultados com os de pesquisas relacionadas que apresentam evidências de concordância e confiabilidade da avaliação por humanos, observa-se que a complexidade do artefato sendo avaliado e seu domínio impacta no grau de concordância e confiabilidade. Em geral, os resultados do presente estudo indicam baixa concordância e boa confiabilidade para alguns subfatores da criatividade do produto. Isso pode estar parcialmente relacionado à complexidade de um artefato de software, quando comparado a artefatos mais simples, criados sob condições experimentais restritas (Tabela 1). Outras pesquisas relacionadas à avaliação de fatores subjetivos, como estética visual de artefatos computacionais, também identificaram discordância sobre um número considerável de interfaces de usuário de aplicativos Android [23], e não atingimento do limite mínimo de concordância entre avaliadores analisando processos de software SPICE [17]. Esses resultados podem indicar uma maior dificuldade para avaliar artefatos complexos com concordância e confiabilidade neste domínio.

A avaliação da criatividade por parte dos professores, requer a capacidade de reconhecer o que é criatividade e se esta está presente ou não no trabalho dos alunos [11]. Conforme apontado por estudos relacionados, professores podem enfrentar dificuldades no reconhecimento da criatividade devido à falta de compreensão de seus subfatores. Considerando a subjetividade da criatividade do produto, os avaliadores nem sempre concordam sobre o que constitui um produto criativo, e a adoção de técnicas com pouca orientação para a avaliação, por exemplo a Técnica de Avaliação Consensual [15], pode levar a uma baixa concordância generalizada entre avaliadores. Portanto, pode ser importante definir detalhadamente o que avaliar, por exemplo, na forma de uma rubrica, a fim de fornecer uma orientação focada visando avaliações mais consistentes e uniformes.

Conforme também apontado por pesquisas relacionadas [40], a experiência dos avaliadores sobre a criatividade impacta a avaliação da criatividade do produto. Enquanto alguns estudos indicam a importância de avaliadores especialistas para obter resultados confiáveis [40], outros reportam que pessoas não

especialistas possuem uma concepção robusta da criatividade [44] e avaliam de forma muito diferente dos especialistas. No que diz respeito à formação dos avaliadores, os resultados indicam que a expertise com relação ao domínio influencia positivamente a confiabilidade das avaliações, portanto, é importante ter professores bem treinados especificamente em computação para fazer uma avaliação com concordância e confiabilidade. No entanto, isso é um desafio para o ensino de computação na Educação Básica, pois atualmente faltam professores formados em licenciatura em computação e com experiência com o ensino de criatividade na educação. Trabalhos relacionados [11] também indicam que os professores podem ser mais eficazes na avaliação da criatividade se tiverem treinamento e experiência relacionada a criatividade. Além disso, professores podem avaliar a criatividade de forma mais eficaz, por exemplo, se tiverem experiência pessoal de trabalho criativo ou se possuírem uma definição robusta do que compõe um produto criativo. Ambas são formas de aprimorar ou refinar as concepções de criatividade dos professores e seu reconhecimento e avaliação da criatividade como um todo.

Em geral, historicamente, grande parte da pesquisa sobre criatividade tem se baseado em julgamentos subjetivos de avaliadores humanos para avaliar aspectos de originalidade, utilidade e condensação. Embora tais abordagens de avaliação manual tenham se mostrado úteis, elas enfrentam duas limitações principais (custo de mão de obra e subjetividade) que, conforme apontado pelos resultados de presente estudo, ameaçam a concordância e confiabilidade. Além disso, em um contexto educacional, os professores podem ter que avaliar um número maior de artefatos criados por uma turma, levando à fadiga e ameaçando ainda mais a confiabilidade [21]. Assim, uma possibilidade para pesquisas futuras é automatizar parte da avaliação da criatividade a fim de apropriar-se do progresso recente na avaliação automatizada para mitigar limitações relacionadas a subjetividade e tempo [61]. Assim, suportando sua aplicação amplamente no ensino de computação na Educação Básica, considerando especificamente a falta de professores bem treinados nesta área de conhecimento [7].

## 5.1 Ameaças à validade

Por ser uma pesquisa exploratória, este estudo apresenta algumas limitações. A fim de analisar as questões de pesquisa no respectivo contexto, foram utilizados exemplos reais de aplicativos criados no contexto do ensino de computação na Educação Básica, diferente de outros estudos que criaram artefatos hipotéticos e muitas vezes simplificados.

Foram utilizadas questões diretas, relacionadas à cada um dos subfatores da criatividade, assumindo que o construto medido é unidimensional, pois ainda não existe um consenso sobre a dimensionalidade, definição e nem escala na literatura. Além disso, foi utilizada uma escala ordinal com apenas 3 pontos para garantir a coleta de dados para todos os pontos da escala, o que também pode ter afetado as respostas. Como no presente estudo, o contexto do aplicativo, a descrição e as capturas de tela foram apresentadas como parte do formulário de avaliação, a maioria dos avaliadores pode ter considerado suficiente para analisar o aplicativo e pode não ter feito o esforço de baixar e experimentar o aplicativo em si. Isso pode ter limitado uma visão mais completa dos aplicativos e, portanto, os resultados da avaliação.

Em termos de significância estatística, obteve-se um tamanho amostral pequeno, porém aceitável para analisar a concordância e confiabilidade entre os avaliadores. A pequena quantidade de participantes restringe a generalização dos resultados. No entanto, mesmo neste contexto limitado, foi observada uma discordância considerável, e espera-se que essa discordância aumente com mais participantes. Além disso, algumas das análises relacionadas à questão 2 de pesquisa podem ter sido impactadas devido às diferenças nos tamanhos das amostras de cada uma das categorias de expertise. Assim, especialmente os casos com pouca representatividade (por exemplo, apenas dois avaliadores da área de Psicologia) requerem pesquisas futuras para confirmar ou não os resultados obtidos. Para minimizar as ameaças relacionadas à análise de dados, foi realizada uma avaliação estatística usando técnicas apropriadas para avaliar a concordância e confiabilidade entre vários participantes para dados em uma escala ordinal. Além disso, foi também realizada uma análise sobre a influência de várias características intraindividuais na avaliação.

# 6 CONCLUSÃO

Com a necessidade de promoção do pensamento criativo e da capacidade de criar soluções inovadoras no contexto do ensino em computação, é importante fornecer métodos de avaliação confiáveis para prover feedback ao aluno e ao professor como parte do processo de aprendizagem. Ainda assim, os resultados do presente estudo indicam níveis muito baixos de concordância e confiabilidade entre avaliadores, sugerindo que a avaliação da criatividade do produto focando em aplicativos móveis como resultados tangíveis de aprendizagem, difere por sua subjetividade inerente e dificuldade de identificar a criatividade do produto e seus subfatores de uma forma consensual. Assim, os resultados demonstram que a concordância entre avaliadores e a confiabilidade da avaliação da criatividade do produto é uma questão que requer uma investigação empírica coordenada. Esses resultados contribuem para discussões sobre métodos de avaliação da criatividade que precisam evoluir na educação em computação.

# AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e pelo CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

# REFERÊNCIAS

[1] A. Aldave, J. M. Vara, D. Granada, E. Marcos, 2019. Leveraging creativity in requirements elicitation within agile software development: a systematic literature review. *Journal of Systems and Software*, 157, 110396.

[2] R. F. Adler, K. Beck, 2020. Developing an Introductory Computer Science Course for Pre-service Teachers. *Journal of Technology and Teacher Education*, 28(3), 519-541.

[3] N. da C. Alves, C. Gresse von Wangenheim, P. E. Rodrigues, J. C. R. Hauck, A. F. Borgatto, 2016. Ensino de Computação de Forma Multidisciplinar em Disciplinas de História no Ensino Fundamental – Um Estudo de Caso. *Revista Brasileira de Informática na Educação*, 24(3).

[4] T. M. Amabile, 1982. Social psychology of creativity: A Consensual Assessment Technique. *Journal of Personality and Social Psychology*, 43, 997-1013.

[5] J. Baer, J. C. Kaufman, C. A. Gentile, 2004. Extension of the Consensual Assessment Technique to Nonparallel Creative Products. *Creativity Research Journal*, 16:1, 113-117.

[6] B. Barbot, R. W. Hass, R. Reiter-Palmon, 2019. Creativity assessment in psychological research: (Re)setting the standards. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2), 233–240.

[7] R. Beaty, D. R. Johnson, 2020. Automating Creativity Assessment with SemDis: An Open Platform for Computing Semantic Distance. *PsyArXiv*. Disponível em: https://doi.org/10.31234/osf.io/nwvps

[8] S. P. Besemer, K. O'Quin, 1986. Analyzing creative products: Refinement and test of a judging instrument. *Journal of Creative Behavior*, 20(2), 115–126.

[9] S. P. Besemer (1998). Creative product analysis matrix: Testing the model structure and a comparison among products—three novel chairs. *Creativity Research Journal*, 11(4), 333–346.

[10] S. Besemer, D. J. Treffinger, (1981). Analysis of Creative Products: Review and Synthesis. Journal of Creative Behavior, 15(3), 158-178.

[11] B. Bolden, C. DeLuca, T. Kukkonen, S. Roy, J. Wearing, 2019. Assessment of Creativity in K-12 Education: A Scoping Review. *Review of Education*.

[12] J. K. Buelin-Biesecker, E. N. Weibe, 2013. Can pedagogical strategies affect students' creativity? Testing a choice-based approach to design and problem-solving in technology, design, and engineering education. *Proc. of the American Society for Engineering Education Annual Conference & Exposition*, Atlanta, GA.

[13] L. J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334.

[14] D. H. Cropley, A. J. Cropley, 2010. Recognizing and fostering creativity in design education. International Journal of Technology and Design Education, 20, 345-358.

[15] G. M. Cseh, K. K. Jeffries, 2019. A scattered CAT: A critical evaluation of the consensual assessment technique for creativity research. Psychology of Aesthetics, Creativity, and the Arts, 13(2), 159–166.

[16] J. Diedrich, M. Benedek, E. Jauk, A. Neubauer, 2015. Are Creative Ideas Novel and Useful?. *Psychology of Aesthetics, Creativity, and the Arts*. 9. 35-40. doi:10.1037/a0038688.

[17] K. El Emam, L. Briand, R. Smith, 1996. Assessor agreement in rating SPICE processes. *Software Process: Improvement and Practice*, 2(4).

[18] J. P. Fatt, 2000. Fostering creativity in education. *Education*, 120 (4), 744–757.

[19] M. N. F. Ferreira, C. Gresse von Wangenheim, R. Missfeldt Filho, F. da Cruz Pinheiro, J. C. R. Hauck, 2019. Learning user interface design and the development of mobile applications in middle school. ACM Interactions, 26(4).

[20] J. L. Fleiss, M. C. Paik, B. Levin. 2003. Statistical Methods for Rates and Proportions. 3rd ed. John Wiley; Sons, Inc.

[21] B. Forthmann, H. Holling, N. Zandi, A. Gerwig, P. Çelik, M. Storme, T. Lubart, 2017. Missing creativity: The effect of cognitive workload on rater (dis-)agreement in subjective divergent-thinking scores, *Thinking Skills and Creativity*, v. 23, pp. 129-139.

[22] M. Graham, A. Milanowski, J. Miller, 2012. Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings. *Center for Education Compensation Reform*. Online Submission.

[23] C. Gresse von Wangenheim, J. V. Araujo Porto, J. C.R. Hauck, A. F. Borgatto, 2018. Do we agree on user interface aesthetics of Android apps? Available at: *arXiv*:1812.09049 [cs.SE], 2018.

[24] H. C. Hill, C. Y. Charalambous, M. A. Kraft, 2012. When rater reliability is not enough: Teacher observation systems and a case for the G-study. *Educational Researcher*, 41(2), p. 56-64.

[25] A. D. Ho, T. J. Kane, 2013. The reliability of classroom observations by school personnel. Research Paper. *MET Project*. Bill & Melinda Gates Foundation.

[26] K. A. Hallgren, 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1).

[27] P. W. Jackson, S. Messick, 1964. The Person, The Product, and the Response: Conceptual Problems in the Assessment of Creativity. *ETS Research Bulletin Series*, 2, i-27.

[28] J. C. Kaufman, J. Baer, J. C. Cole, J. D. Sexton, 2008. A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, 20 (2), 171–178. doi:10.1080/10400410802059929

[29] J. C. Kaufman, J. Baer, 2005. The amusement park theory of creativity. In J. C. Kaufman & J. Baer (Eds.), *Creativity across domains: Faces of the muse*, 321-328. Hillsdale, NJ: Lawrence Erlbaum Associates.

[30] J. C. Kaufman, J. Baer (2012). Beyond new and appropriate: who decides what is creative? *Creativity Research Journal*, 24, 83–91. doi:10.1080/10400419.2012.649237

[31] J. C. Kaufman, Lauren E. Skidmore, 2010. Taking the Propulsion Model of Creative Contributions into the 21st Century. *Psychologie in Österreich*, 5.

[32] S. A. Kornilov, T. V. Kornilova, E. L. Grigorenko, 2016. The cross-cultural invariance of creative cognition: A case study of creative writing in U.S. and Russian college students. *New Directions for Child and Adolescent Development*, 151 (2016), 47-59.

[33] R. S. Kramer, M. Mileva, K. L. Ritchie, 2018. Inter-rater agreement in trait judgements from faces. *PloS one*, 13(8), e0202655.

[34] K. Krippendorff, 2016. Misunderstanding reliability. *Methodology*, 12(4), 139-144.

[35] D. J. Krus, G. C. Helmstadter, 1993. The problem of negative reliabilities. *Educational and Psychological Measurement*, 53, 643-650.

[36] I. Lee, F. Martin, J. Denner, B. Coulter, W. Allan, J. Erickson, J. Malyn-Smith, L. Werner, 2011. Computational thinking for youth in practice. *ACM Inroads* 2(1)(March 2011), 32–37.

[37] S. C. Liao, E. A. Hunt, W. Chen, 2010. Comparison between inter-rater reliability and inter-rater agreement in performance assessment. Annals Academy of Medicine Singapore, 39(8), 613.

[38] R. Litchfield, 2008. Brainstorming Reconsidered: A Goal-Based View. *The Academy of Management Review*, 33, 649-668.

[39] R. Mohanani, P. Ram, A. Lasisi, P. Ralph and B. Turhan, 2017. Perceptions of Creativity in Software Engineering Research and Practice In: 43rd *Euromicro Conference on Software Engineering and Advanced Applications* (SEAA), Vienna, 2017, 210-217. doi: 10.1109/SEAA.2017.21.

[40] D. R. Mullet, A. Willerson, K. N. Lamb, T. Kettler, 2016. Examining teacher perceptions of creativity: A systematic review of the literature. *Thinking Skills and Creativity*, 21, 9-30.

[41] J. D. Novak, 2013. Meaningful learning is the foundation for creativity. *Qurriculum-Revista de Teoría, Investigación y Práctica Educativa*, 26, 27-38.

[42] K. O'Quin, 1987. Creative product scale: applications in product improvement. In Creativity and Innovation: towards a European Network. *Report of the First European Conference on Creativity and Innovation, 'Network in Action'*, organized by the Netherlands Organization for Applied Scientific Research TNO Delft, The Netherlands.

[43] M. Resnick, K. Robinson. 2017. Lifelong kindergarten: Cultivating creativity through projects, passion, peers, and play. MIT press.

[44] J. Pétervári, 2018. The evaluation of creative ideas–analysing the differences between expert and novice judges (Doctoral dissertation, Queen Mary University of London).

[45] S. M. Reis, J. S. Renzulli, 1991. The assessment of creative products in programs for gifted and talented students. Gifted Child Quarterly, 35(3), 128–134.

[46] M. Rhodes, 1961. An Analysis of Creativity. The Phi Delta Kappan, 42(7), 305-310.

[47] M. Romero, A. Lepage, B. Lille, 2017. Computational thinking development through creative programming in higher education. International Journal of Educational Technology in Higher Education, 14(1), 42.

[48] S. Said-Metwaly, E. Kyndt, W. Van den Noortgate, 2017. Approaches to Measuring Creativity: A Systematic Literature Review. Creativity - Theories - Research - Applications, 4(2).

[49] C. L. Semmelroth, E. Johnson, 2014. Measuring inter rater reliability on a special education observation tool. Assessment for Effective Intervention, 39(9), 131-145.

[50] J. S. Smith, 2016. Assessing Creativity: Creating a Rubric to Effectively Evaluate Mediated Digital Portfolios. *Journalism & Mass Communication Educator*, 72(1), 24-36.

[51] R. J. Sternberg, 2012. The assessment of creativity: An investment-based approach Creativity Research Journal, 24 (2012), 3-12.

[52] R. J. Sternberg, 2018. What's wrong with creativity testing?. The Journal of Creative Behavior, 54(1), 20-36.

[53] R. J. Sternberg, J. Kaufman, J. Pretz, 2004. A Propulsion Model of Creative Leadership. *Creativity and Innovation Management*, 13, 145-153.

[54] I. A. Taylor, B. J. Sandler, 1972. Use of a creative product inventory for evaluating products of chemists. In Proceedings of the 80th *Annual Convention of the American Psychological Association*, 7, 311–312.

[55] M. Tissenbaum, J. Sheldon, H. Abelson, 2019. From Computational Thinking to Computational Action. *Comm. of the ACM*, 62(3), 34-36.

[56] S. M. Todd, S. Shinzato, 1999. Thinking for the future: Developing higher level thinking and creativity for students in Japan—and elsewhere. Childhood Education, 75 (6), 342–345. doi:10.1080/00094056.1999.10522054

[57] C. Walia, 2019. A Dynamic Definition of Creativity. Creativity Research Journal, 31(3), 237-247.

[58] K. L. Westberg, 1996. The Effects of Teaching Students How to Invent. The Journal of Creative Behavior, 30, 249-267.

[59] A. Yadav, S. Cooper, 2017. Fostering Creativity Through Computing. Comm. of the ACM, 60(2), 31-33.

[60] J. Zhou, X. Wang, J. Song, J. Wu, 2017. Is It New? Personal and Contextual Influences on Perceptions of Novelty and Creativity. *Journal of Applied Psychology*, 102.

[61] R. A. Beghetto. 2019. Large-Scale Assessments, Personalized Learning, and Creativity: Paradoxes and Possibilities. ECNU Review of Education, 2(3), 311-327.