Proposta de Escala para Avaliação da Aprendizagem de *Machine Learning* em nível *Create* na Educação Básica

Marcelo Fernando Rauber[1, 2], Christiane Gresse von Wangenheim[1], Adriano F. Borgatto[1], Ramon Mayor Martins[1], Deise M. Arndt[1], Jean Carlo Rossa Hauck[1]

[1]Universidade Federal de Santa Catarina (UFSC) - Florianópolis - SC - Brasil.
[2]Instituto Federal Catarinense (IFC) - Camboriú - SC - Brasil.

marcelo.rauber@ifc.edu.br, c.wangenheim@ufsc.br,
adriano.borgatto@ufsc.br, ramon.mayor@posgrad.ufsc.br,
deise.arndt@posgrad.ufsc.br, jean.hauck@ufsc.br

Abstract. *There is a trend to include the teaching of Machine Learning (ML) in K-12, teaching students to create their own intelligent solutions. In this context, it is also important to assess the students' learning of ML and the Design Thinking process at the Create level. Evolving an assessment model, the aim of this article is to propose a scale with pedagogical interpretation using Item Response Theory (IRT). The results provide a first indication of the suitability of the assessment model in terms of internal consistency and IRT calibration parameters that are very close to acceptable. We expect the definition of the scale to support the learning of the creation of ML solutions by providing feedback to students and teachers.*

Resumo. *Há uma tendência de incluir o ensino de Machine Learning (ML) já na Educação Básica, ensinando os alunos a criar suas próprias soluções inteligentes. Nesse contexto, também é importante avaliar o aprendizado de ML e do processo de Design Thinking em nível Create. Evoluindo um modelo de avaliação, este artigo tem como objetivo propor uma escala com sua interpretação pedagógica utilizando a Teoria de Resposta ao Item (TRI). Os resultados fornecem uma primeira indicação da adequação do modelo de avaliação em relação à consistência interna e parâmetros de calibração da TRI muito próximos aos aceitáveis. Esperamos que a definição da escala possa apoiar o aprendizado da criação de soluções de ML fornecendo feedback aos alunos e professores.*

1. Introdução

Aplicações de Inteligência Artificial (IA) e *Machine Learning* (ML) permeiam nossas vidas (Unesco, 2022). No entanto, muitas pessoas não entendem a tecnologia envolvida (House of Lords, 2018). Portanto, torna-se importante introduzir conceitos e práticas básicas de ML já na escola (Camada e Durães, 2020; Caruso e Cavalheiro, 2021), permitindo que os alunos se tornem não apenas consumidores conscientes de aplicativos de ML, mas também criadores de soluções inteligentes e eticamente corretas (Kandlhofer *et al*., 2016; Royal Society, 2017). Seguindo as diretrizes curriculares (Touretzky *et al*., 2019; Long e Magerko, 2020) e também conforme indicado pela Base Nacional Comum Curricular para o Ensino Médio (MEC, 2022), se deve incluir uma compreensão dos conceitos básicos de ML, como algoritmos de aprendizagem e fundamentos de redes neurais, bem como limitações e considerações éticas relacionadas ao ML. Para atingir esse objetivo, várias iniciativas estão surgindo, adotando principalmente metodologias ativas que orientam os alunos a aprender como preparar um conjunto de dados, treinar o modelo de ML, avaliar seu desempenho e usá-lo para

DOI: 10.5753/sbie.2024.241471

realizar previsões (Martins e Gresse von Wangenheim, 2023) usando tipicamente ferramentas visuais, como o Google *Teachable Machine* (Google, 2023). Mesmo que a maioria dos cursos atuais de ML tenham como público-alvo alunos iniciantes, alguns começam a enfocar no estágio *Create* do ciclo *Use-Modify-Create* (Lee *et al*., 2011), incentivando os estudantes a desenvolver seus próprios projetos de ML, levando a soluções *open-ended* nas quais não há um produto final pré-definido, o que permite explorar possibilidades infinitas e não se limitar a uma única resposta ou solução.

Ainda se observa uma carência de abordagens para a avaliação da aprendizagem de conceitos ML em nível *Create* com confiabilidade e validade (Rauber e Gresse von Wangenheim, 2022). De forma geral, as propostas existentes para avaliar a aprendizagem de ML na Educação Básica são escassas e relativamente simples, usando tipicamente quizzes ou autoavaliações (Rauber e Gresse von Wangenheim, 2022). Uma forma de conduzir a avaliação da aprendizagem é por meio da avaliação baseada em desempenho, que analisa artefatos criados pelos alunos como resultado da aprendizagem, tipicamente suportada por uma rubrica de pontuação (Moskal, 2000; Morrison *et al*., 2019), que permite determinar notas a partir dos níveis de desempenho. Exemplos de modelos de avaliação de aprendizagem de ML com base no desempenho incluem a rubrica em nível *Use* proposta por Gresse von Wangenheim *et al*. (2021), posteriormente avaliada com boa consistência interna (Ômega global = 0,834) e qualidade dos itens quanto a discriminação e diferenciação (Rauber *et al*., 2023), porém limitada à avaliação de um produto final pré-definido. Existe também uma proposta inicial de uma rubrica em nível *Create*, avaliada por um painel de especialistas, que aponta para adequação do modelo no contexto da Educação Básica, com concordância moderada entre os especialistas (*inter-rater*) na aplicação prática da rubrica (Fleiss Kappa = 0,534) e validade do conteúdo dos itens em termos de correção, relevância, integridade e clareza (Rauber e Gresse von Wangenheim, 2023).

O cálculo das notas das rubricas pode variar, usando desde métodos simples como porcentagens ou soma de pontuações dos itens das rubricas até a criação de escalas com a Teoria de Resposta ao Item (TRI) (Mislevy *et al.*, 2003; Mislevy, 2012). Uma escala pode ser definida como o nível em que as variáveis teóricas não observáveis diretamente podem ser inferidas a partir de um instrumento de medição composto por um conjunto de itens (DeVellis, 2017). Muitas vezes tida como superior à soma de pontuações dos itens , a TRI é "*uma coleção de modelos de medição que visam explicar as conexões entre as respostas observadas dos itens em uma escala e um construto subjacente*" (Cappelleri, Lundy, e Hays, 2014), permitindo a criação de uma escala, na qual tanto o nível de dificuldade de cada item quanto o nível de habilidade do aluno podem ser inferidos (DeVellis, 2017; Paek e Cole, 2020).

Assim, este artigo tem como objetivo criar e interpretar uma escala, a partir de uma rubrica em nível *Create* (Rauber e Gresse von Wangenheim, 2023; Rauber *et al*., 2023), utilizando como metodologia de análise a TRI (DeVellis, 2017; Paek e Cole, 2020), a fim de evoluir o modelo de avaliação baseado em desempenho da aprendizagem de ML.

2. Modelo de avaliação baseado em desempenho em nível *Create*

Uma alternativa para ensinar ML em nível *Create* para alunos nos Anos Finais do Ensino Fundamental e Médio é o curso de curta duração "Apps inteligentes para Todos!" (Almeida, 2022). Esse curso é voltado ao nível *Create*, ensinando o estudante a identificar uma necessidade, criar seu próprio modelo de ML para classificação de imagens e implantá-lo em um aplicativo móvel com MIT App Inventor. Seguindo o processo de *Design Thinking* (Brown, 2008), o curso leva o aluno a identificar um problema em relação a sua vida cotidiana ou comunidade, que possa ser resolvido utilizando classificação de imagens. A partir deste problema, o aluno é estimulado a propor uma solução de ML útil e usável, desenvolver um modelo de ML e implementar a solução em um aplicativo móvel. O Google Teachable Machine é usado para treinar e avaliar o desempenho do modelo de ML, que ao final é exportado na Google Cloud. Em seguida, os alunos esboçam as interfaces de usuário em papel, elaboram o design visual e codificam um aplicativo móvel funcional usando o MIT App Inventor (MIT, 2024) utilizando a extensão TMIC que possibilita implantar o modelo de ML desenvolvido (Oliveira *et al*., 2022). O curso está disponível online gratuitamente em português do Brasil em https://cursos.computacaonaescola.ufsc.br/.

Como parte do curso "Apps inteligentes para Todos!" foi projetado um modelo de avaliação baseado em desempenho em nível *Create*, desenvolvido e avaliado seguindo a metodologia de Design Centrado em Evidências (Mislevy *et al*., 2003; Seeratan e Mislevy, 2008). A proposta de Rauber e Gresse von Wangenheim (2023) apresenta um modelo de avaliação que contém uma rubrica para avaliação de desempenho de aprendizagem de computação por meio de desenvolvimento de aplicativos inteligentes por alunos dos Anos Finais do Ensino Fundamental e Médio. O modelo completo inclui a avaliação de vários conceitos e práticas de computação conforme apresentado na Tabela 1.

Tabela 1. Modelo de avaliação de aprendizagem (nível *Create*).

Conceitos e práticas	Rubrica	Confiabilidade e validade
Machine Learning	ML-Create (Rauber e Gresse von Wangenheim, 2023)	Concordância moderada entre avaliadores Fleiss Kappa = 0,534 e validades de face e conteúdo
Design Thinking	DT-Create (Rauber e Gresse von Wangenheim, 2023)	--
Algoritmos e programação	Fluência de programação (Alves *et al*., 2020)	Confiabilidade boa (Cronbach α = 0,84) e com validade convergente
Design de interface de usuário e estética	Conformidade do *design* com guias de estilo (Solecki *et al*., 2020)	Confiabilidade boa (Cronbach α = 0,84) e com validade convergente e discriminante
	Modelo *Deep Learning* para analisar a estética visual de telas (Lima, 2023)	Validade concorrente com uma correlação de Spearman de ρ = 0.9
Criatividade	Originalidade, flexibilidade e fluência de aplicações móveis (Alves, 2023)	Confiabilidade boa (Ômega = 0.86) com validade de construto e dos itens

Se observa que o modelo de avaliação integra várias rubricas já existentes e validadas anteriormente, apontando a sua confiabilidade e validade. Assim, o foco do presente artigo é voltado às rubricas DT-Create para avaliação do processo de *Design Thinking* (Tabela 2) e ML-Create voltadas ao desenvolvimento do modelo de ML (Tabela 3). Essas rubricas, hora revisadas e padronizadas para quatro níveis de desempenho, foram inicialmente avaliadas por um painel de especialistas, cujos resultados apontam para adequação do modelo no contexto da Educação Básica, com concordância estatisticamente moderada entre os especialistas (*inter-rater*) na aplicação

prática da rubrica (Fleiss Kappa = 0,534) e validade do conteúdo dos itens em termos de correção, relevância, integridade e clareza (Rauber e Gresse von Wangenheim, 2023).

Tabela 2. Rubrica de avaliação de desempenho de aprendizagem de *Design Thinking*

ID	Critério / Variáveis observáveis	Níveis de Desempenho			
		Não entregue - 0 pontos	Fraco - 1 ponto	Aceitável - 2 pontos	Bom - 3 pontos
Descoberta					
I01	Descrição do problema geral	Não descreveu	Descreveu incompleto e/ou incorreto	-	Descreveu completo e correto
I02	Descrição do público alvo	Não descreveu	Descreveu incompleto e/ou incorreto	-	Descreveu completo e correto
I03	Descrição do ambiente de uso	Não descreveu	Descreveu incompleto e/ou incorreto	-	Descreveu completo e correto
Ideação					
I04	Descrição geral da solução	Não descreveu	Descreveu incompleto e/ou incorreto	-	Descreveu completo e correto
I05	Especificação dos requisitos funcionais (Exceto de ML)	Não descreveu	Descreveu incompleto e/ou incorreto	-	Descreveu completo e correto
I06	Especificação dos requisitos de usabilidade	Não descreveu	Descreveu incompleto e/ou incorreto	-	Descreveu completo e correto
Testes					
I07	Teste Funcional	Não descreveu	Descreveu incompleto e/ou incorreto	-	Descreveu completo e correto
I08	Teste de usabilidade	Não descreveu	Descreveu incompleto e/ou incorreto	-	Descreveu completo e correto
Compartilhamento					
I09	Compartilhamento	Não descreveu	Compartilhou o caderno ou o software inteligente em ao menos uma rede social.	-	Compartilhou o caderno ou o software inteligente em duas ou mais redes sociais.

3. Metodologia de Pesquisa

A fim de evoluir o modelo de avaliação baseado em desempenho (Rauber e Gresse von Wangenheim, 2023; Rauber *et al*., 2023), este estudo tem como objetivo criar uma escala e sua interpretação pedagógica com base na análise estatística usando a TRI (DeVellis, 2017; Paek e Cole, 2020), das competências de ML relacionadas à classificação de imagens e do processo de *Design Thinking* em nível *Create*, por alunos dos Anos Finais do Ensino Fundamental e Médio. Desta forma, são analisadas as seguintes questões de análise: QA1. Há evidências de validade das rubricas no nível *Create* por meio da análise fatorial? QA2. Há evidências da confiabilidade das rubricas por meio da TRI? QA3. É possível criar uma interpretação pedagógica das escalas segundo o resultado da TRI?

A pesquisa foi conduzida de forma exploratória[1], com base em um estudo de caso, a partir de dados coletados na aplicação prática do curso “Apps inteligentes para Todos!”. O curso foi aplicado entre outubro e dezembro do ano de 2023 como um programa educacional suplementar para estudantes dos anos finais do Ensino Fundamental e Médio em situação de vulnerabilidade social por uma parceria da iniciativa Computação na Escola da Universidade Federal de Santa Catarina (CnE, 2024) com o Programa PodeCrer do Instituto Vilson Groh (IVG, 2024). Foi utilizada uma amostragem não-probabilística aplicando o método de amostragem de conveniência. Um total de 105 estudantes foram matriculados no curso, formando seis turmas no contra turno escolar. Destes, 77 alunos entregaram todos os artefatos avaliados, com idades entre 14 e 21 anos (dos quais 6 têm mais de 18 anos), sendo 39 do sexo feminino e 38 do sexo masculino atribuído ao nascimento. O curso foi ofertado com 16 horas/aula por turma ao longo de 8 semanas, com aulas semanais de duas horas, no laboratório de informática do Instituto Vilson Groh (Figura 1-A). Foi ministrado por

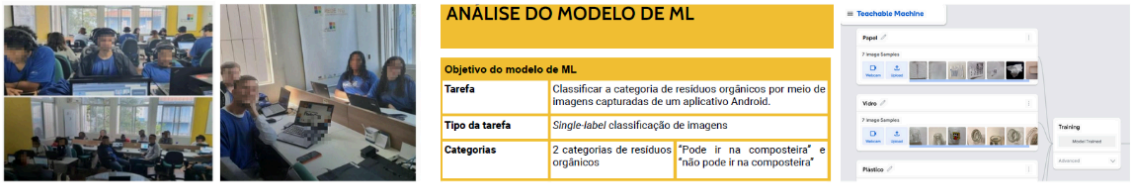
[1] Este estudo foi aprovado pelo Comitê de Ética da Universidade Federal de Santa Catarina (Pareceres nº 4.893.560 e nº 5.610.912)

dois doutorandos em Ciência da Computação, com mais de dez anos de experiência no ensino de computação para o público-alvo, especializados nas áreas de IA/ML, contando com suporte de monitores. Os instrutores forneceram orientação sobre como realizar as atividades e responderam a todas as perguntas, enquanto os alunos seguiram instruções passo a passo fornecidas como material on-line.

Tabela 3. Rubrica de avaliação de desempenho de aprendizagem de ML – nível *Create*

ID	Critério / Variáveis observáveis	Níveis de Desempenho			
		Não entregue - 0 pontos	Fraco - 1 ponto	Aceitável - 2 pontos	Bom - 3 pontos
Análise de requisitos de ML					
I10	Objetivo do modelo de ML especificado	Não descreveu	Descrição incompleta ou incorreta	-	Descrição completa e correta
I11	Especificação de riscos e requisitos de desempenho	Não descreveu	Descrição incompleta ou incorreta	Descrição parcialmente completa/correta	Descrição completa e correta
I12	Análise dos riscos de erro	Não descreveu	Incorreta identificação do nível de risco	-	Descrição completa e correta
I13	Acurácia correta em relação ao risco de erro	Não descreveu	Descrição incompleta ou incorreta	-	Descrição completa e correta
Gerenciamento de dados					
I14	Quantidade de imagens	Não enviou informações (do arquivo .tm)	Menos de 20 imagens por categoria	21 - 35 imagens por categoria	Mais de 36 imagens por categoria
I15	Distribuição do conjunto de dados	Não enviou informações (do arquivo .tm)	A quantidade de imagens em cada categoria varia muito. Mais de 10% de variação em ao menos uma categoria (relativo ao total)	A quantidade de imagens entre as categorias têm entre 3% e 10% de variação	Todas as categorias têm a mesma quantidade de imagens (menos de 3% de variação)
I16	Imagens com conteúdo ético (20% das imagens analisadas)	Não enviou informações (do arquivo .tm)	Ao menos 3 imagens contém conteúdo não ético (violência, nudez, armas)	Ao menos uma imagem não contém imagens não éticas.	Todas as imagens são éticas.
Treinamento do modelo					
I17	Treinamento (Epochs: 50, batch size: 16, Learning rate: 0,001)	Não enviou informações (do arquivo .tm)	O modelo não foi treinado	O modelo foi treinado usando os parâmetros padrões	O modelo foi treinado com parâmetros ajustados
Interpretação de desempenho					
I18	Acurácia mínima (Planejado x Atingido)	Não enviou as informações para viabilizar a análise	A precisão do modelo é baixa, sendo inferior a 10% da precisão mínima planejada	A precisão do modelo está próxima, ficando até 10% abaixo da precisão mínima planejada	O modelo atinge a acurácia mínima planejada
I19	Análise da Acurácia	Não enviou as informações para viabilizar a análise	Análise incorreta da acurácia total modelo	-	Análise e identificação correta da acurácia total do modelo
I20	Interpretação da Acurácia	Nenhuma informação enviada sobre acurácia e/ou interpretação	Incorreta interpretação da análise da acurácia do modelo	-	Correta interpretação da análise da acurácia do modelo
I21	Análise da matriz de confusão	Não enviou as informações para viabilizar a análise	Mais de dois erros na identificação de classificações errôneas	Até dois erros na identificação de classificações errôneas	Identificação correta de erros de classificação
I22	Interpretação da matriz de confusão	Não enviou as informações para viabilizar a análise	Interpretação a respeito do modelo é incorreta	-	Correta interpretação em respeito ao modelo
I23	Ajustes / Melhorias realizadas	Não enviou as informações para viabilizar a análise	Nenhuma nova iteração de desenvolvimento foi relatada	Uma nova iteração com mudanças no conjunto de dados e/ou parâmetros de treinamento foi relatada	Várias iterações com mudanças no conjunto de dados e/ou parâmetros de treinamento foram relatadas
I24	Testes com novos objetos	Não enviou as informações para viabilizar a análise	Nenhum objeto testado	1-3 objetos testados	Mais de 3 objetos testados
I25	Análise dos resultados de testes	Não enviou as informações para viabilizar a análise	Indicação errada da quantidade de erros nos testes	-	Indicação correta da quantidade de erros nos testes
I26	Interpretação dos resultados dos testes	Não enviou as informações para viabilizar a análise	Interpretação errada dos resultados dos testes	-	Correta interpretação dos resultados dos testes

Foram coletados artefatos elaborados pelos alunos como resultados de aprendizagem por meio da plataforma Moodle. Isso inclui o modelo de ML desenvolvido, contido no arquivo (.tm) da Google *Teachable Machine,* e um relatório on-line pré definido preenchido ao longo das aulas, que documenta o processo de *Design Thinking*, bem como o planejamento, análise, interpretação do desempenho e predição do modelo de ML criado para classificação de imagens (Figura 1-B), chamado de Diário de Bordo. Adicionalmente também foi coletado do código do aplicativo (.aia) desenvolvido no MIT App Inventor.



A) Estudantes durante o curso B) Exemplos de artefatos usados na avaliação da aprendizagem

Figura 1. Aplicação e exemplos de artefatos criados

Os artefatos entregues pelos estudantes foram automaticamente avaliados usando as rubricas do modelo de avaliação implementado como parte do CodeMaster (Gresse von Wangenheim *et al.*, 2018), disponível gratuitamente em português do Brasil, on-line, em http://apps.computacaonaescola.ufsc.br/codemaster/. A avaliação de ML e *Design Thinking* se baseia especificamente no modelo de evidência (Tabelas 2 e 3) resultando nas frequências dos níveis de desempenho alcançados pelos alunos apresentados na Tabela 4.

Tabela 4. Distribuição da frequência alcançada por critério de avaliação

Níveis de	Traço Latente Design Thinking									Traço Latente ML nível *Create*																
Desempenho	I01	I02	I03	I04	I05	I06	I07	I08	I09	I10	I11	I12	I13	I14	I15	I16	I17	I18	I19	I20	I21	I22	I23	I24	I25	I26
Não entregue	3	4	15	8	2	4	18	22	1	1	1	0	0	0	0	0	0	0	0	0	12	10	16	0	16	17
Fraco	62	12	52	63	9	9	12	0	39	50	42	0	10	74	15	90	0	9	6	6	2	14	9	1	28	7
Aceitável	-	-	-	-	-	-	-	-	-	-	14	-	-	18	38	0	84	5	0	0	28	0	44	16	0	0
Bom	12	61	10	6	66	64	47	55	37	26	20	77	67	2	41	4	10	63	71	71	35	53	8	60	33	53
Total	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77	77

3.1. Análise dos dados

Foi realizada a preparação típica de dados para análise estatística com a TRI (Bennett e von Davier, 2017; Paek e Cole, 2020), recodificando itens para garantir que as categorias começam em zero e cresçam sequencialmente com mesmo intervalo. Itens com baixa variabilidade nos níveis de desempenho (menor que 5), foram agrupados no nível de desempenho mais baixo. Já os itens I12 e I16 foram desconsiderados na análise dos dados, pois não apresentaram variabilidade nessa amostra. Todos os dados coletados foram reunidos em uma única amostra para análise. Dada a natureza dos dados, composta por itens com respostas politômicas categóricas ordinais, em todas as análises foi adotado o Modelo de Resposta Gradual (MRG) da TRI (Samejima, 1969; Samejima, 1997), utilizando a escala padrão do modelo (0,1), em que zero indica a média do grupo e 1 equivale ao desvio padrão (Paek e Cole, 2019). Para análise dos dados, foi utilizada a linguagem R (R Core Team, 2022), com os pacotes mirt (Chalmers, 2012) e psych (Revelle, 2022). Quanto à questão de análise 1, a validade de um construto se refere à capacidade que os critérios do instrumento conseguem medir o traço latente que o mesmo se propõe a medir (DeVellis, 2017), envolvendo a validade convergente obtida

pelo grau de correlação entre os critérios do instrumento. Primeiro, verificamos a adequação dos dados para a análise fatorial por meio do teste de Kaiser-Meyer-Olkin (KMO) (Brown, 2015). Em seguida, realizamos uma análise fatorial exploratória que permite descobrir a estrutura essencial das variáveis de observação multivariadas e lidar com a redução da dimensionalidade (DeVellis, 2017). Em conjunto com os resultados da carga fatorial, se considerou a qualidade da calibração dos parâmetros ao MRG unidimensional da TRI sem especificações de distribuições a *priori* no modelo (DeVellis, 2017; Paek e Cole, 2019). Itens considerados inadequados foram excluídos nas análises dessa amostra. Quanto à questão de análise 2, a confiabilidade refere-se à estabilidade ou consistência das pontuações dos critérios do instrumento de avaliação em um mesmo fator (Moskal e Leydens, 2000), assim investigamos qualidade do processo de estimação dos parâmetros do modelo da TRI aplicado a rubrica. A qualidade da obtenção das estimativas dos parâmetros foi verificada pelos valores de dificuldade e discriminação, com o objetivo de estabelecer os níveis de dificuldade dos itens e como eles diferenciam entre os diferentes níveis de habilidade dos alunos (DeVellis, 2017; Paek e Cole, 2019). Também foram analisadas as curvas características das categorias dos itens. Com relação à questão de análise 3, nosso objetivo é criar uma interpretação descritiva do desempenho do aluno que está sendo avaliado de acordo com o modelo de avaliação estabelecido (*in press*) usando a TRI seguindo o *Smoothing Method of Scale Anchoring* (Beaton e Allen, 1992). Com base no fato de que a dificuldade dos itens e a habilidade dos alunos (Θ, Theta) são expressas na mesma unidade de acordo com a TRI (DeVellis, 2017; Paek e Cole, 2019), o *Smoothing Method of Scale Anchoring* procura diferenciar o ponto de ancoragem, ou intervalos de valores das habilidades dos alunos , nos quais há maior probabilidade de inferir um determinado nível na escala, com base nas curvas características da categoria de cada item, tentando generalizá-los e descrevê-los com base na descrição de cada nível de desempenho do modelo de avaliação (Beaton e Allen, 1992). Entretanto, no exemplo dado por Beaton e Allen (1992), os itens são dicotômicos e envolvem a probabilidade de um acerto casual, e o ponto de inflexão da curva está em torno de 65%. No nosso caso, com os itens politômicos, cujo ponto de inflexão da curva está em 50%, portanto, o ponto em que o item está posicionado foi adaptado. Além disso, para itens politômicos, todas as categorias são posicionadas.

4. Resultados e Discussão

4.1. Há evidências de validade das rubricas no nível *Create* por meio da carga fatorial da análise fatorial?

Iniciamos conduzindo o teste de KMO, no qual tipicamente valores acima de 0,5 são assumidos como adequados para realização de uma análise fatorial útil (Brown, 2015; DeVellis, 2017). Analisando os itens da rubrica, obtivemos um índice KMO de 0,58 para os itens do traço latente de *Design Thinking* e um KMO de 0,49 para os itens do traço latente de ML nível *Create*, respectivamente adequado e muito próximo à condição de aceitabilidade, indicando que os resultados da análise fatorial devem ser interpretados com cuidado. Conduzimos a análise fatorial exploratória separadamente para cada traço latente. Paralelamente, consideramos a qualidade da calibração dos itens na TRI. No resultado da calibração da TRI, o parâmetro "a" indica o padrão de

discriminação e está associado à qualidade do item e a carga fatorial, enquanto o valor de "b_i" pode ser interpretado como indicadores relativos de dificuldade, indicando o limiar de passagem de um nível de desempenho inferior para um superior do item (DeVellis, 2017; Paek e Cole, 2019). A Tabela 5 apresenta os resultados iniciais encontrados para ambos os traços latentes.

Tabela 5. Carga fatorial e qualidade (iniciais) da calibração da TRI

			Carga Fatorial	Discri-minação	Dificuldade		
			F1	a	b_1	b_2	b_3
Traço Latente 1 Design Thinking	I01	Descrição do problema geral	0,575	1,196	1,763	NA	NA
	I02	Descrição do público alvo	0,409	0,762	-1,959	NA	NA
	I03	Descrição do ambiente de uso	0,622	1,350	-1,387	1,824	NA
	I04	Descrição geral da solução	**0,119**	**0,204**	**-10,622**	**12,203**	NA
	I05	Especificação dos requisitos funcionais (Exceto de ML)	0,977	**7,730**	-1,086	NA	NA
	I06	Especificação dos requisitos de usabilidade	0,950	5,168	-1,005	NA	NA
	I07	Teste Funcional	**0,084**	**0,144**	**-8,369**	-3,199	NA
	I08	Teste de usabilidade	**-0,009**	**-0,015**	**62,650**	NA	NA
	I09	Compartilhamento	**-0,015**	**-0,025**	-3,146	NA	NA
Traço Latente 2 ML-Create	I10	Objetivo do modelo de ML especificado	**-0,139**	**-0,239**	-2,813	NA	NA
	I11	Especificação de riscos e requisitos de desempenho	**-0,089**	**-0,152**	-1,460	**-6,841**	NA
	I13	Acurácia correta em relação ao risco de erro	**-0,041**	**-0,071**	**27,041**	NA	NA
	I14	Quantidade de imagens	**-0,194**	**-0,337**	-3,568	NA	NA
	I15	Distribuição do conjunto de dados	**-0,177**	**-0,306**	**6,002**	-1,088	NA
	I17	Treinamento	0,637	1,405	1,811	NA	NA
	I18	Acurácia mínima (Planejado x Atingido)	**0,115**	**0,197**	**-10,273**	**-7,635**	NA
	I19	Análise da Acurácia	**-0,109**	**-0,186**	**13,387**	NA	NA
	I20	Interpretação da Acurácia	**-0,140**	**-0,241**	**10,380**	NA	NA
	I21	Análise da Matriz de Confusão	0,336	**0,608**	-2,990	0,315	NA
	I22	Interpretação da Matriz de Confusão	**0,125**	**0,215**	**-8,883**	-3,684	NA
	I23	Ajustes e melhorias realizados	0,447	0,849	-1,731	-0,914	2,854
	I24	Testes com novos objetos	**-0,532**	-1,070	1,470	NA	NA
	I25	Análise dos resultados dos testes	0,999	**40,355**	-0,713	0,209	NA
	I26	Interpretação dos resultados dos testes	0,961	**5,890**	-0,712	-0,405	NA

Na análise da qualidade dos itens do traço latente de *Design Thinking*, os itens I04, I07, I08 e I09 são candidatos a serem desconsiderados, pois todos apresentam carga fatorial abaixo do recomendado de 0,3 (Brown, 2015) e também com baixo padrão de discriminação, abaixo de 0,7 (De Ayala, 2022; DeVellis, 2017). Ao mesmo tempo, os itens I04, I07 e I08 apresentam valores discrepantes para os limiares de dificuldade, que se espera estarem entre -5 e +5 (DeVellis, 2017; Paek e Cole, 2019). Excluídos esses itens, chegamos a resultados adequados (vide seção 3.2). Já a qualidade dos itens do traço latente de ML em nível *Create*, há vários candidatos a exclusões. Foi executada uma sequência de operações, sempre excluindo um item por vez e calibrado novamente os modelos para acompanhar os resultados. O item I25 foi o primeiro a ser considerado candidato a ser desconsiderado, por apresentar uma carga fatorial extremamente alta e ao mesmo tempo com um padrão de discriminação alto. Calibrando novamente o modelo, em seguida o item I19 e depois I22 foram excluídos pelos mesmos motivos. Realizadas essas exclusões, chegamos a resultados muito próximos do adequado.

4.2 Há evidências da confiabilidade das rubricas por meio da TRI?

Em geral, se observou bons resultados de validade e confiabilidade (Tabela 6). Com relação ao traço latente de *Design Thinking* todos os itens apresentaram carga fatorial

adequada, variando entre 4,41 (I02) e 0,777 (I05). Já com relação ao traço latente de ML em nível *Create*, o menor valor foi de 0,26 (I15) e o maior foi de 0,421 (I20). Mesmo assim, o menor valor é próximo do considerado adequado (0,3). O que evidencia a validade convergente das rubricas, isto é, a capacidade que os critérios do instrumento conseguem medir o traço latente a que se propõe. Com relação à qualidade da calibração dos itens no modelo GRM da TRI, de forma geral os resultados são adequados. O traço latente de *Design Thinking* apresenta discriminação dentro de condições de aceitabilidade, bem como os indicadores de dificuldade estão em condições de aceitabilidade. Já o traço latente de ML em nível *Create* apresenta itens com padrão de discriminação ligeiramente abaixo do desejável (I10, I11, I12, I13, I14, I15, I18, I21, I23, I24 e I25) mas mesmo assim aceitáveis, dado o tamanho da amostra. Mesmo o item com menor padrão de discriminação, I15, pode ser considerado aceitável já que seus indicadores de dificuldade estão dentro do aceitável e a carga fatorial bem próxima do considerado adequado. Quanto aos indicadores de dificuldade do traço latente de ML nível *Create*, estes estão adequados. De forma geral, e especialmente ao considerarmos o tamanho da amostra, há indicativos que esses primeiros resultados podem ser aceitos.

Tabela 6. Carga fatorial e qualidade da calibração do modelo GRM da TRI

			Carga Fatorial	Discriminação	Dificuldade		
			F1	a	b1	b2	b3
Traço Latente 1 Design Thinking	I01	Descrição do problema geral	0,460	0,883	2,219	NA	NA
	I02	Descrição do público alvo	0,410	0,765	-1,978	NA	NA
	I03	Descrição do ambiente de uso	0,551	1,123	-1,593	2,096	NA
	I05	Especificação dos requisitos funcionais (Exceto de ML)	0,777	2,103	-1,450	NA	NA
	I06	Especificação dos requisitos de usabilidade	0,776	2,095	-1,295	NA	NA
Traço Latente 2 ML nível Create	I10	Objetivo do modelo de ML especificado	0,357	**0,650**	1,121	NA	NA
	I11	Especificação de riscos e requisitos de desempenho	0,372	**0,682**	0,421	1,709	NA
	I13	Acurácia correta em relação ao risco de erro	0,336	**0,608**	-3,316	NA	NA
	I14	Quantidade de imagens	0,323	**0,580**	2,174	NA	NA
	I15	Distribuição do conjunto de dados	**0,260**	**0,459**	-4,016	0,751	NA
	I17	Treinamento	0,380	0,700	2,933	NA	NA
	I18	Acurácia mínima (Planejado x Atingido)	0,356	**0,648**	-3,320	-2,493	NA
	I20	Interpretação da Acurácia	0,421	0,789	-3,407	NA	NA
	I21	Análise da Matriz de Confusão	0,311	**0,557**	-3,203	0,310	NA
	I23	Ajustes e melhorias realizados	0,378	**0,695**	-2,085	-1,150	3,320
	I24	Testes com novos objetos	0,318	**0,572**	-2,339	NA	NA
	I26	Interpretação dos resultados dos testes	0,352	**0,639**	-2,122	-1,349	NA

Outra forma de visualizar a relação entre a habilidade dos estudantes, a discriminação e a dificuldade dos itens usando a TRI com itens politômicos ordinais, são as curvas características de categoria (CCC), que indicam a probabilidade de selecionar uma opção categórica para cada item como uma função do nível de habilidade do aluno (Figura 2). Theta (Θ) representa o nível de habilidade estimado para um aluno, enquanto P(Θ) é a probabilidade de escolher a opção correspondente com referência a um determinado valor de Θ. Para os itens com apenas dois níveis de desempenho após os ajustes da preparação típica dos dados com TRI (itens I01, I02, I05, I06, I10, I13, I14, I17, I20 e I24), apenas uma curva é plotada, representando o limiar entre as duas categorias remanescentes. É importante observar a adequação de todas as categorias definidas para cada item no modelo de avaliação. Ao analisar as CCCs das respostas de dos itens com duas categorias de respostas (apenas uma linha plotada), coerentemente o aumento da probabilidade do aluno ter sua avaliação

classificada no nível de desempenho mais alto em cada item está relacionado com o aumento da habilidade do aluno representada pelo Theta. Já os itens com mais de duas categorias de resposta, todas as CCCs mostram picos sucessivos para cada opção de resposta e estão ordenadas como seria de se esperar, inclusive a categoria "Não entregue". Além disso, o ponto máximo da curva para a categoria mais baixa (“Não entregue”) está à esquerda de todas as outras categorias, indicando que um valor Theta mais baixo leva a uma maior probabilidade dessa categoria ser inferida ao longo do *continuum* Theta, conforme esperado (DeVellis, 2017; Paek e Cole, 2019). Da mesma forma, a categoria mais alta (“Aceitável” ou “Bom”) de cada item está mais à direita das outras, indicando que um valor Theta mais alto leva à uma probabilidade maior dessa categoria ser inferida ao longo do *continuum* Theta, conforme esperado (DeVellis, 2017; Paek e Cole, 2019).

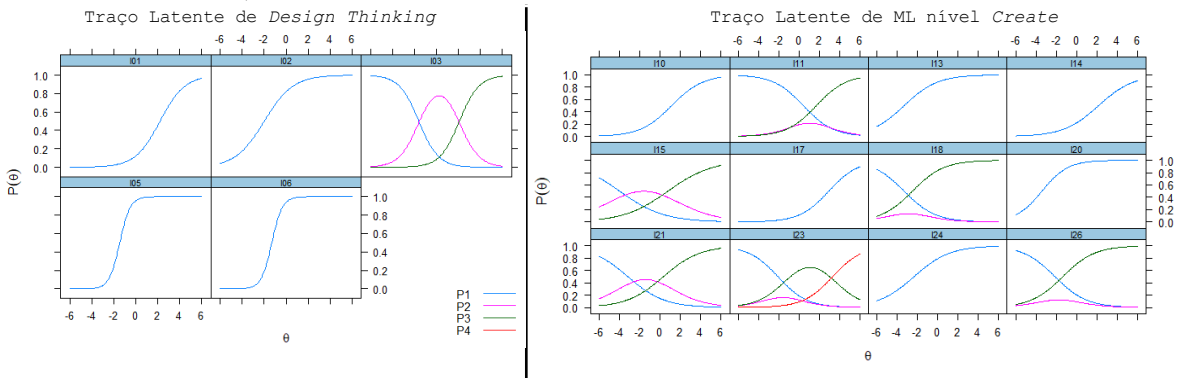


Figura 2. Curvas características de categoria para cada item

Analisando a discriminação das CCCs, é possível observar que a inclinação das linhas no ponto médio indica seu bom padrão de discriminação. Ela também representa adequadamente as habilidades dos alunos, uma vez que os baixos níveis de habilidade dos alunos (Theta) tendem a apresentar uma probabilidade maior de atingir níveis de desempenho mais baixos. Ao mesmo tempo, os níveis crescentes de habilidade (Theta) tendem a uma probabilidade maior de atingir níveis de desempenho mais altos e coerentemente sequenciais. Isso reflete a boa aderência e os valores de ajuste adequados do modelo TRI usado para as características latentes.

Algumas CCCs são mais achatadas e até mesmo baixas das outras curvas do mesmo item, indicando uma probabilidade menor de atingir o nível de desempenho correspondente. Normalmente, as categorias com um pico de curva de probabilidade abaixo de 10% devem ser revisadas ou agrupadas com outras. Os resultados mostram que mesmo as categorias com linhas de probabilidade abaixo das outras são adequadas, pois apresentam picos de curva de probabilidade aceitáveis (I11-Aceitável = 21,60%, I18-Aceitável = 13,41%, I23-Fraco = 16,10%, I26-Fraco = 12,28%). Até mesmo a curva achatada de ambos os traços latentes, "I26-Interpretação dos resultados dos testes", na categoria "Fraco", pode ser considerada adequada, tendo um pico de curva com 12,28% de probabilidade de ser atingido (próximo a Theta -1,8), o que demonstra sua relevância, não devendo ser desconsiderada.

4.3 É possível criar uma interpretação pedagógica das escalas segundo a TRI?

Adotamos o *Smoothing Method of Scale Anchoring* (Beaton e Allen, 1992) para inferir a dificuldade de cada categoria de resposta dos itens. De maneira análoga, é possível

alocar a faixa de dispersão inerente à probabilidade de cada categoria de item a ser inferida com base no desempenho de um aluno na escala proposta (Figura 3).

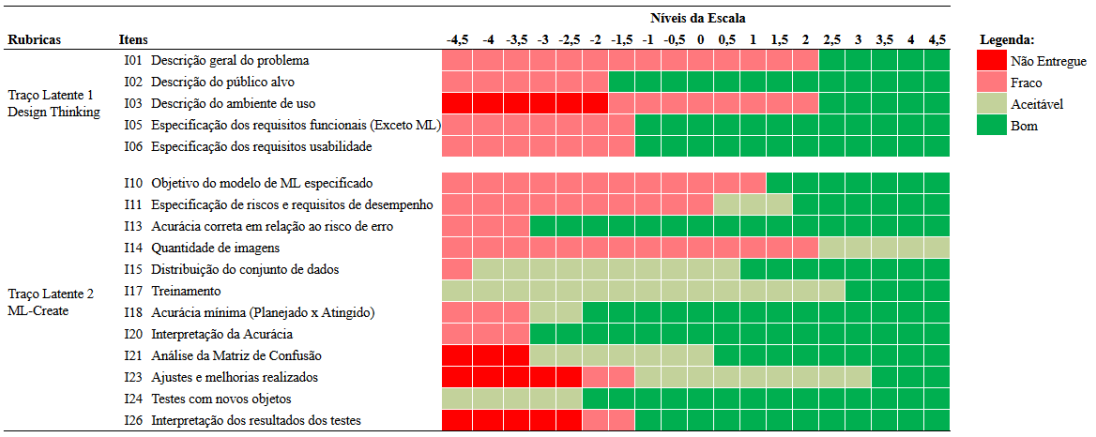


Figura 3. Faixa da dispersão da probabilidade do nível de desempenho dos alunos

Assim, pode ser inferida uma interpretação educacional a ser dada às pontuações obtidas pelos alunos (Tabela 7). Isso é possível porque, uma vez que a escala tenha sido especificada, ela pode ser interpretada no contexto do aprendizado de ML, em que o nível de dificuldade de cada categoria de item para a escala estabelecida e os intervalos com referência a um determinado Theta podem ser logicamente relacionados (Beaton e Allen, 1992). A análise realizada do traço latente é cumulativa, ou seja, à medida que o traço latente aumenta, ele abrange as habilidades descritas nos níveis anteriores.

4.4 Ameaças à validade

A fim de minimizar impactos de validade neste estudo, identificamos ameaças potenciais e aplicamos estratégias de mitigação. Para atenuar a ameaça da validade da conclusão, adotamos uma metodologia sistemática, definindo claramente o objetivo do estudo, a coleta de dados e a análise estatística. Os métodos estatísticos foram selecionados com cuidado, seguindo o procedimento proposto por DeVellis (2017) para a construção de escalas de medição e com os procedimentos típicos da TRI (Bennett e von Davier, 2017; Paek e Cole, 2020), e assim, antes de prosseguir com a calibração do modelo da TRI, realizamos a análise fatorial e analisamos a qualidade da calibração, que se mostrou adequado ou muito próximo. Outra questão refere-se à qualidade dos dados agrupados em uma única amostra. Isso foi possível pela padronização dos dados, todos coletados da mesma maneira em aplicações do curso “Apps inteligentes para Todos!” para o mesmo tipo de público alvo. Outro risco se refere à validade das pontuações alocadas com base nos dados coletados. Para os campos textuais e descritivos foi considerado o tamanho da resposta (contagem de palavras). Mas esse risco é minimizado, já que a automação se mostrou adequada diante da análise de conteúdo manual de cinco alunos com rendimentos distintos. Não foram consideradas imagens inseridas no Diário de Bordo. Entretanto, este risco é minimizado, pois as análises foram realizadas de forma automatizada (utilizando um script Python dentro da ferramenta CodeMaster), a partir das mesmas rubricas. Outra ameaça à validade externa está associada ao tamanho da amostra e à diversidade dos dados utilizados. Nossa análise é baseada em uma amostra de 77 alunos. Isto é considerado um tamanho de amostra suficiente para uma pesquisa exploratória, porém levando em consideração os

resultados das análises, a amostra deve ser aumentada no futuro para revisar os resultados obtidos.

Tabela 7. Interpretação descritiva do desempenho dos alunos

Rubrica	Θ Theta value	Análise descritiva
Traço Latente 1 Design Thinking	menor que -1,0	O estudante consegue minimamente seguir as instruções para desenvolver as etapas de Design Thinking. Ele apresenta uma descrição deficitária do problema, do público alvo, da especificação dos requisitos funcionais e dos requisitos de usabilidade. Dificilmente percebe a relevância da descrição do ambiente de uso.
	de -1,0 até menor que -2,5	O estudante consegue seguir as instruções para desenvolver as etapas de Design Thinking. Descreve adequadamente o público alvo, bem como especifica os requisitos funcionais e requisitos de usabilidade. Apresenta uma descrição deficitária do problema e do ambiente de uso.
	Maior ou igual a 2,5	O estudante consegue seguir as instruções para desenvolver as etapas de Design Thinking e obtém as melhores notas em todos os critérios considerados. Todos os aspectos são adequadamente descritos, incluindo a descrição geral do problema, do público alvo, do ambiente de uso e a especificação dos requisitos funcionais e requisitos de usabilidade.
Traço Latente 2 Machine Learning - nível Create	Menor que -3,0	O estudante consegue minimamente seguir as instruções para criar, treinar seu modelo de ML com os parâmetros padrões e fazer alguns testes com novas imagens. Utiliza poucas imagens por categoria (menos de 20) mas a distribuição das imagens nas categorias têm uma variação aceitável (de 3% a 10% de variação). Apresenta uma descrição deficitária do objetivo do modelo de ML, dos riscos e requisitos de desempenho, e da acurácia em relação ao risco de erro. A acurácia atingida é bem inferior à planejada (mais de 10%), e tem dificuldade para interpretar a acurácia. Não consegue realizar a análise da matriz de confusão, e também não percebe a importância de realizar ajustes/melhorias e não consegue interpretar os resultados dos testes.
	de -3,0 até menor que 0,5	O estudante consegue seguir as instruções para criar, treinar seu modelo de ML com os parâmetros padrões e realizar testes com novas imagens, interpretando os resultados desses testes corretamente. Utiliza poucas imagens por categoria (menos de 20) mas a distribuição das imagens nas categorias têm uma variação aceitável (de 3% a 10% de variação). Apresenta uma descrição deficitária do objetivo do modelo de ML e dos riscos e requisitos de desempenho. Define corretamente a acurácia em relação ao risco de erro, bem como, a interpreta corretamente. O modelo de ML criado atinge a acurácia planejada. Tem noção de como realizar a análise da matriz de confusão mas ainda comete erros. Relata ter tentado ao menos uma vez realizar ajustes/melhorias no modelo de ML.
	maior ou igual 0,5	Obtém as melhores notas na maioria dos critérios considerados. O estudante consegue seguir as instruções para criar, treinar seu modelo de ML e até alterar os parâmetros padrões, e, realizar testes com novas imagens, interpretando os resultados desses testes corretamente. Utiliza poucas imagens por categoria (menos de 35) e a distribuição das imagens nas categorias têm uma pequena variação (menos de 3%). Apesar de poder encontrar dificuldade, apresenta uma boa descrição objetivo do modelo de ML e dos riscos e requisitos de desempenho. Define corretamente a acurácia em relação ao risco de erro, bem como, a interpreta corretamente. O modelo de ML criado atinge a acurácia planejada. Realiza corretamente a análise da matriz de confusão. Relata ter tentado realizar ajustes/melhorias no modelo de ML.

5. Conclusão

De modo geral, os resultados dessa avaliação mostram que é possível criar e interpretar uma escala para a avaliação baseada no desempenho das competências de ML e de *Design Thinking* em nível *Create*, relacionadas à classificação de imagens, sob a perspectiva de pesquisadores no contexto educacional dos anos finais do Ensino Fundamental e Médio, aplicando o curso “Apps inteligentes para Todos!” na prática. Ambos os traços latentes mostraram validade convergente dos construtos, demonstrado pela análise fatorial exploratória em conjunto com a qualidade dos parâmetros de calibração. Ambos os traços latentes também mostraram confiabilidade, evidenciada pela consistência dos parâmetros discriminação e dificuldade dos itens de cada traço latente. Como resultado principal, foi possível criar uma interpretação descritiva do desempenho dos alunos, que deve ser considerada em conjunto para avaliação da aprendizagem. Esses resultados e a implementação realizada (em conjunto com as demais rubricas como parte do modelo de avaliação) têm o potencial de auxiliar em um processo de avaliação tanto ao fornecer feedback aos estudantes quanto à avaliação da sua aprendizagem. Com esses primeiros resultados positivos, é importante que trabalhos futuros repliquem o estudo com amostras maiores e em outros contextos visando confirmar os achados.

Agradecimentos

Gostaríamos de agradecer a toda equipe do IVG e a todos os estudantes que participaram dos cursos.

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil (CNPq).

Referências

Almeida, B. C. S. (2022). *Desenvolvimento de um Curso Ensinando a Criação de Apps Inteligentes para a Classificação de Imagens com Machine Learning e Design Thinking*. Trabalho de conclusão de curso, Graduação em Sistemas de Informação/ UFSC, Brasil.

Alves, N. C., Gresse von Wangenheim, C., Hauck, J. C. R., and Borgatto, A. F. (2020). A Large-scale Evaluation of a Rubric for the Automatic Assessment of Algorithms and Programming Concepts. *In Proc. of ACM Technical Symposium on Computer Science Education*. Portland, OR, USA.

Alves, N. C. (2023). *Assessing the Creativity of Mobile Applications in Computing Education*. PhD Thesis, PPGCC/UFSC, Brazil.

Beaton, A., and Allen, N. (1992), Interpreting Scales Through Scale Anchoring. *Journal of Educational Statistics*, 17(2).

Bennett, R. E., and von Davier, M. (2017). *Advancing human assessment: The methodological, psychological and policy contributions of ETS*. Switzerland: Springer Nature.

Brown, T. (2008). Design thinking. *Harvard business review*, 86(6).

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Second edition. New York, USA: The Guilford Press.

Camada M. Y. and Durães G. M., (2020). Ensino da Inteligência Artificial na Educação Básica: um novo horizonte para as pesquisas brasileiras. *Proc. of Simpósio Brasileiro de Informática na Educação*, *online*, Brazil.

Cappelleri, J. C., Jason Lundy, J., and Hays R. D. (2014). Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures. *Clinical Therapeutics*, 36(5).

Caruso A. L. M. and Cavalheiro S. A. da C., (2021). Integração entre Pensamento Computacional e Inteligência Artificial: uma Revisão Sistemática de Literatura. *Proc. of Simpósio Brasileiro de Informática na Educação*, *online*, Brazil.

Chalmers, Robert P. (2012), mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*. 48(6).

CnE. (2024), *Computação na Escola*. Retrieved 22/05/2024 from https://computacaonaescola.ufsc.br/

De Ayala, R. J. (2022), *The Theory and Practice of Item Response Theory*. Second edition. Guilford Press, New York, NY, USA.

DeVellis, R. F. (2017), *Scale development: theory and applications.* Fourth edition. Los Angeles, USA: SAGE.

Google (2023). Google Teachable Machine. Retrieved 01/06/2023 from https://teachablemachine.withgoogle.com/.

Gresse von Wangenheim, C. G. von, Hauck, J. C. R., Demetrio, M. F., Pelle, R., Cruz Alves, N. da, Barbosa, H. and Azevedo, L. F. (2018), CodeMaster—Automatic Assessment and Grading of App Inventor and Snap! Programs. *Informatics in Education*, 17(1).

Gresse von Wangenheim C., Alves N. da C., Rauber M. F., Hauck J. C. R., and Yeter I. H. (2021). A Proposal for Performance-based Assessment of the Learning of Machine Learning Concepts and Practices in K-12. *Informatics in Education,* 21(3).

House of Lords (2018), *AI in the UK: ready, willing and able*. HL Paper 100, London, UK.

IVG. (2023). *Instituto Pe. Vilson Groh*. Retrieved 22/05/2024 from https://redeivg.org.br/

Kandlhofer, M., Steinbauer, G., Hirschmugl-Gaisch, S., and Huber, P. (2016), Artificial intelligence and computer science in education: From kindergarten to university. *Proc. of the Frontiers in Education Conference*, Erie, PA, USA, 1–9.

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

Lima, A. L. S. (2023). *Automated assessment of the visual aesthetics of App Inventor user interfaces with Deep Learning*. PhD Thesis, PPGCC/UFSC, Brazil.

Long, D. and Magerko, B. (2020), What is AI literacy? Competencies and design considerations. *Proc. of the Conf. on Human Factors in Computing Systems*, Honolulu, HI, USA.

Martins, R. M. and Gresse von Wangenheim, C, (2023). Findings on Teaching Machine Learning in High School: A Ten - Year Systematic Literature Review. *Informatics in Education*, 22 (3).

Mislevy R. J., Almond R. G., and Lukas J. F., (2003), A Brief Introduction to Evidence-Centered Design. *ETS Research Report Series*, 2003(1).

Mislevy, R. J. (2012), Design and Discovery in Educational Assessment: Evidence-Centered Design, Psychometrics, and Educational Data Mining. Design and Discovery in Educational Assessment: Evidence-Centered Design. *Psychometrics, and Educational Data Mining*, 4(1).

Morrison, G. R., Ross, S. M., Morrison, J. R., and Kalman, H. K., (2019), *Designing effective instruction*. Eighth editioned. Hoboken, NJ: Wiley.

Moskal B. M. and Leydens J. A., (2000), Scoring rubric development: Validity and reliability. *Practical assessment, research, and evaluation*, 7(1).

MEC (2022), Normas sobre Computação na Educação Básica – Complemento à Base Nacional Comum Curricular (BNCC). Parecer 02/2022 CNE/CEB/MEC.

Oliveira, F. P., Gresse von Wangenheim, C., and Hauck, J. C. R. (2022). TMIC: App Inventor Extension for the Deployment of Image Classification Models Exported from Teachable Machine. arXiv:2208.1263

Paek, I., and Cole, K. (2020), *Using R for Item Response Theory Model Applications*. New York, NY, USA: Routledge.

R Core Team. (2022), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rauber M. F. and Gresse Von Wangenheim C., (2022), Assessing the Learning of Machine Learning in K-12: A Ten-Year Systematic Mapping. *Informatics in Education*, 22(2).

Rauber, M. F., Gresse von Wangenheim, C., Barbetta, P. A., Borgatto, A. F., Martins, R. M. and Hauck, J. R. (2023). Reliability and Validity of an Automated Model for Assessing the Learning of Machine Learning in Middle and High School: Experiences from the "ML for All!" course. *Informatics in Education*, online.

Rauber, M. F. and Gresse von Wangenheim, C., (2023), Uma proposta para avaliação do desempenho de aprendizagem de conceitos e práticas de Machine Learning em nível Create na Educação Básica. *In Proc. of Simpósio Brasileiro de Informática na Educação*, SBC, Passo Fundo, Brazil.

Revelle, W. (2022), *psych: Procedures for Personality and Psychological Research*. Northwestern University, Evanston, USA. https://CRAN.R-project.org/package=psych Version = 2.2.3.

Royal Society, (2017), Machine learning: the power and promise of computers that learn by example. Retrieved 01/06/2022 from royalsociety.org/machine-learning.

Samejima, F. (1969), Estimation of latent ability using a response of graded scores. Monograph 17. *Psychometrika*, 34(2).

Samejima, F. (1997), Graded response model. *Handbook of Modern Item Response Theory*. New York, NY, USA: Springer.

Seeratan, K. L., and Mislevy, R. J. (2008), *Design patterns for assessing internal knowledge representations*. Menlo Park, USA: SRI International.

Solecki, I., Porto, J., Alves, N. D. C., Gresse von Wangenheim, C., Hauck, J., and Borgatto, A. F. (2020). Automated Assessment of the Visual Design of Android Apps Developed with App Inventor. *In Proc. of ACM Technical Symposium on Computer Science Education*. Portland, OR, USA.

Touretzky, D., Gardner-McCune, C., Martin, F., and Seehorn D. (2019). Envisioning AI for K-12: What Should Every Child Know about AI? *Proc. of the AAAI Conference on Artificial Intelligence*, Honolulu, HI, USA.

Trochim, W. M. K., and Donnelly, J. P. (2008). *The research methods knowledge base*. Third edition. Mason: Atomic Dog/Cengage Learning.

UNESCO (2022). *K-12 AI curricula: a mapping of government-endorsed AI curricula.* Retrieved 06/06/2022 from https://unesdoc.unesco.org/ark:/48223/pf0000380602