Informatics in Education, 2024, Vol. 23, No. 2, 409–437
© 2024 *Vilnius University*
DOI: 10.15388/infedu.2024.10

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

Marcelo Fernando RAUBER[1,2],
Christiane GRESSE VON WANGENHEIM[1],
Pedro Alberto BARBETTA[3], Adriano FERRETI BORGATTO[3],
Ramon Mayor MARTINS[1], Jean Carlo Rossa HAUCK[1]

[1]*Graduate Program in Computer Science, Department of Informatics and Statistics, Federal University of Santa Catarina, Florianópolis/SC, Brazil.*
[2]*Federal Institute Catarinense (IFC), Camboriú/SC, Brazil.*
[3]*Graduate Program in Methods and Management in Evaluation, Federal University of Santa Catarina, Florianópolis/SC, Brazil.*
e-mail: marcelo.rauber@ifc.edu.br, c.wangenheim@ufsc.br, pedro.barbetta@ufsc.br, adriano.borgatto@ufsc.br, ramon.mayor@posgrad.ufsc.br, jean.hauck@ufsc.br

Received: April 2023

Abstract. The insertion of Machine Learning (ML) in everyday life demonstrates the importance of popularizing an understanding of ML already in school. Accompanying this trend arises the need to assess the students' learning. Yet, so far, few assessments have been proposed, most lacking an evaluation. Therefore, we evaluate the reliability and validity of an automated assessment of the students' learning of an image classification model created as a learning outcome of the "ML for All!" course. Results based on data collected from 240 students indicate that the assessment can be considered reliable (coefficient Omega = 0.834/Cronbach's alpha α = 0.83). We also identified moderate to strong convergent and discriminant validity based on the polychoric correlation matrix. Factor analyses indicate two underlying factors "Data Management and Model Training" and "Performance Interpretation", completing each other. These results can guide the improvement of assessments, as well as the decision on the application of this model in order to support ML education as part of a comprehensive assessment.

Keywords: K-12, middle and high school, Machine Learning, Artificial Intelligence, neural network, image classification, assessment, evaluation.

1. Introduction

Our culture, diversity, education, scientific knowledge, communication, and information are deeply impacted by a diverse set of Artificial Intelligence (AI) technologies, with Machine Learning (ML) being one of the most prominent fields (UNESCO, 2022). ML refers to systems that learn and evolve from their own experience without having to be explicitly programmed to do some task, based on a mathematical/statistical model from data (Mitchell, 1997). More recently, Deep Learning approaches using neural networks, provide substantial progress in ML improving the state of art, for example, in computer vision through image recognition/classification (LeCun *et al.*, 2015; UNESCO, 2022).

However, a significant portion of the population does not understand the technology used in ML, which can make it mysterious or even scary (Ho and Scadding, 2019). To demystify what ML is, how it works, and demonstrate its impacts and limitations, there is a growing need for the public to understand ML (House of Lords, 2018). Thus, it is important to introduce basic ML concepts already at school (Camada and Durães, 2020; Caruso and Cavalheiro, 2021), guiding students to critically and consciously use ML models, as well as providing a first contact to create intelligent and ethically correct solutions (Kandlhofer *et al.*, 2016; Royal Society, 2017; UNESCO, 2022).

Following the curriculum guidelines proposed by Touretzky *et al.* (2019) and Long and Magerko (2020), teaching ML should start at K-12 and cover an understanding of basic ML concepts, such as learning algorithms and neural network fundamentals, as well as limitations and ethical considerations related to ML. Aiming to achieve this goal, several initiatives are emerging proposing the teaching of AI/ML in K-12 (Long and Magerko, 2020; Marques *et al.*, 2020). Expecting students to go beyond merely understanding ML concepts, but to becoming creators of ML models, typically active learning methodologies are adopted with a focus on the human-centered development of an ML model (Amershi *et al.*, 2019), in order to teach students how to prepare a dataset, train a ML model, and evaluate its performance and use it for the prediction of new images (Lwakatare *et al.*, 2019; Ramos *et al.*, 2020). Therefore, typically visual no-coding tools are used, such as Google Teachable Machine (GTM; Google, 2023). This allows students to run an ML process interactively, through a cycle of training, feedback, and correction, enabling them to evaluate the performance of the ML model and so, making changes to the model aiming to improve the model (Gresse von Wangenheim *et al.*, 2021). Adopting the Use-Modify-Create cycle (Lee *et al.*, 2011) commonly novices are taught to first inspect and manipulate pre-defined ML models on the Use stage, then to modify these models until at the Create stage, students are encouraged to develop their own ML projects.

As part of this learning process, it is important to assess the student's performance and to provide feedback to both the student and instructor (Hattie and Timperley, 2007). Assessment as the result of an educational experience, comprising both the process of collecting and analyzing information from various sources, aims to understand in depth the student's knowledge, what s/he understands, and what tasks s/he can accomplish (Huba and Freed, 2000). Yet, so far few models have been proposed to assess the learning of ML in K-12 ranging from quizzes and self-assessments to performance-based models assessing the learning of basic ML concepts, approaches, and some cases ethical issues and the impact of ML on lower cognitive levels (Rauber and Gresse von Wangenheim, 2022). Most of these assessments are done manually by the instructor with only very few automated solutions. Furthermore, a lack of evaluation of these assessments is common, leaving their reliability and validity questionable (Marques *et al.*, 2020, Rauber and Gresse von Wangenheim, 2022). A few exceptions include Hsu *et al.* (2022) analyzing the reliability of a five-item self-assessment questionnaire reporting a Cronbach's alpha $\alpha = 0.883$, and Hitron *et al.* (2019) analyzing the reliability of the coding performed by researchers when manually labeling student's short answer items of an essay on basic ML understanding reporting an interrater Kappa of 92%. Only Shamir and Levin (2021) report the analysis of content validity, in which students and instructors reviewed the questions for the ability to read and understand the items, but without providing statistical results.

Therefore, this research aims at evaluating a performance-based assessment model of the learning of concepts and practices regarding image classification with artificial neural networks in middle and high school proposed by Gresse von Wangenheim *et al.* (2021) based on data collected from the application of the "ML for All!" course (Gresse von Wangenheim *et al.*, 2020). The assessment is based on the examination of student-created artifacts as a part of open-ended applications on the Use stage of the Use-Modify-Create cycle (Lee *et al.*, 2011). An initial evaluation of the scoring rubric through an expert panel demonstrates its internal consistency as well as its correctness and relevance. In order to evolve the evaluation of the proposed assessment model, we analyze the reliability and validity of the assessment based on data collected from 240 middle and high school students.

2. Assessment of Learning ML Concepts at the Use Stage in Middle and High School

Aiming to teach ML to middle and high school students, an alternative is the course "ML for All!" (Gresse von Wangenheim *et al.*, 2020) presenting basic concepts of ML and artificial neural networks as well as teaching students to develop a predefined model for the classification of recycling thrash images. As part of the course ML for All! (Gresse von Wangenheim *et al.*, 2020), we systematically designed, developed, and implemented an assessment model adopting the Evidence-Centered Design methodology (Mislevy *et al.*, 2003).

Domain Analysis. The target audience are Brazilian students from public middle and high schools, with a minimum age of 12 years. At this educational stage, it is expected that students are fluent in their native language, have developed logical-mathematical reasoning, and know how to perform everyday activities with computers, such as accessing the Internet (MEC, 2018). Yet, as computing education in Brazil in practice is still limited to extracurricular programs (Santos *et al.*, 2018) many students do not have computing competencies nor AI/ML knowledge.

Regarding access to computer resources, 66% of urban school students have at least one computer or tablet at home, and 42% spend more than three hours a day dealing with technology (TIC Educação, 2019). Furthermore, 98% of students in urban schools have access to the Internet via a smartphone.

In Brazil, most schools do not have teachers with a degree in computing education (MEC, 2020). Therefore, computing education is typically introduced in an interdisciplinary way being taught by teachers from diverse backgrounds in curricular knowledge areas, such as history, science, etc., with few computing competencies. This can complicate the assessment of learning outcomes and even result in unreliable results. Furthermore, as public school classes are typically large with sometimes more than 30 students, a manual assessment of the students' projects represents an effort and time consuming activity.

Following the K-12 Guidelines for Artificial Intelligence (Touretzky *et al.*, 2019a) referring to the Big Idea 3 – Learning, AI literacy (Long and Magerko, 2020) and a human-centered ML process (Amershi *et al.*, 2019) the course "ML for All!" (Gresse von Wangenheim *et al.*, 2020) aims to promote knowledge building regarding basic ML concepts with a focus on image recognition including data preparation, model training, performance evaluation and prediction of ML model at the Use stage. In order to operationalize the teaching of ML with an interdisciplinary approach related to the Sustainable Development Goals (United Nations, 2015), the course focuses on the task of the classification of recyclable trash images.

Domain Modeling. Based on the domain analysis, the Principled Assessment Designs for Inquiry design pattern were adopted, specifying the elements that will be needed in the assessment considering computational artifacts in the context of computing education of ML (Mislevy *et al.,* 2003; Seeratan and Mislevy, 2008).

a) **Student competencies.** The student model is designed considering focal knowledge, skills, and abilities (Mislevy and Haertel, 2006) as well as other knowledge/skills/abilities that may be required. At the Use stage of the "Use-Modify-Create" cycle (Lee *et al.*, 2011), following a human-centered ML process (Amershi *et al.*, 2019) as well as AI curricula guidelines, the general learning objective is to introduce students to Machine Learning enabling them to have a basic understanding of how ML and neural networks work and develop an ML model for image classification. Table 1 presents the learning objectives with respect to ML competencies at the Use stage in the course "ML for All!".

b) **Task model.** The task model designs a family of potential tasks, and how to structure the kinds of situations we need to obtain the kinds of evidence needed for the evidence models (Mislevy *et al.*, 2003). Elements for the task model are described in Table 2.

   The assessment is defined to be applied as part of the course "ML for All!" (Gresse von Wangenheim *et al.*, 2020) teaching ML to middle and high school students without prior knowledge of computing or AI/ML. The course is planned to be taught in eight hours. The course teaches basic concepts on ML and neural networks and how to develop a predefined ML model for image recognition following the basic steps of a human-centric ML process including data preparation, model training, performance evaluation, and prediction. Aiming at an

Table 1

Student model for the assessment of ML competencies at Use stage

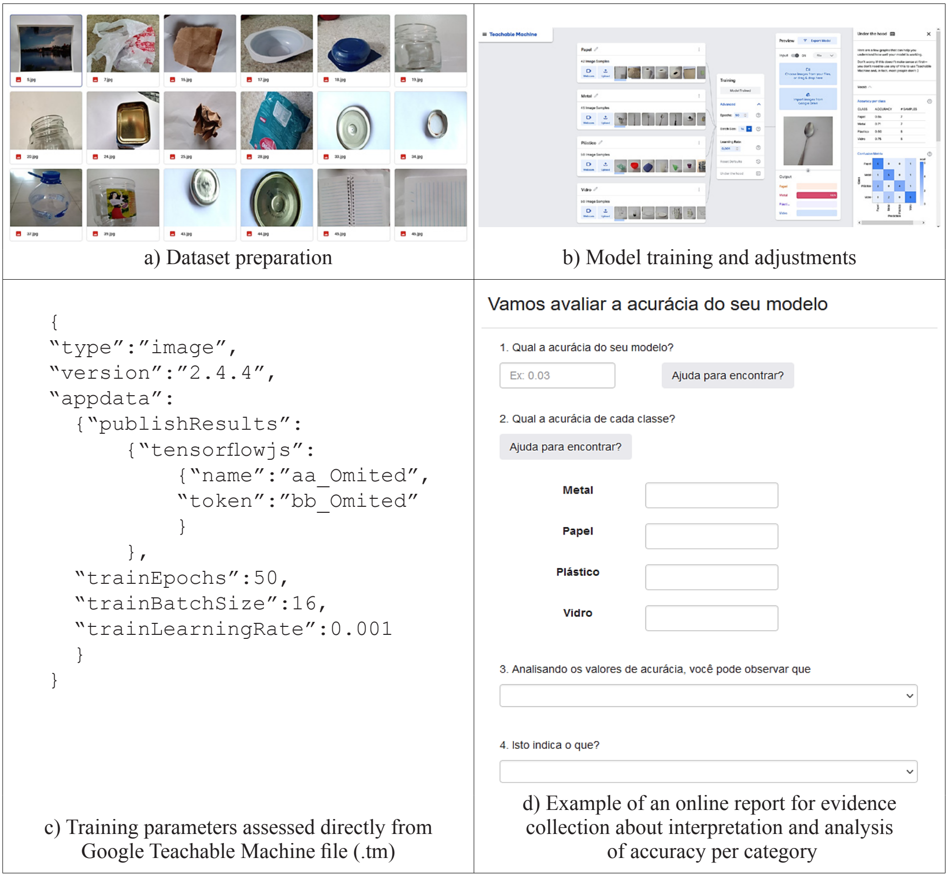
Element	Description
Rationale	Neural networks are a current and key technique in ML for the development of image classification models.
Focal knowledge, skills, and other attributes	Understanding of basic concepts about neural networks. Ability to collect, clean and label data for the training of an ML model. Ability to train an ML model for image classification using a visual tool. Ability to analyze and interpret the performance of the trained ML model and to improve the model. Ability to test the ML model with new images for prediction.
Additional knowledge, skills, and attributes	Ability and maturity to understand instructions in Brazilian Portuguese. The ability to use a computer (basic operations) and access the Internet via a browser. Ability to login to web pages with personal user data.

Table 2

Task model for the assessment of ML competencies at Use stage

Element	Description
Potential work products	Google Teachable Machine file (.tm) including the dataset and category labels. Report on the evaluation of the test results with new objects. Report on the evaluation of the model's performance (accuracy table and confusion matrix). Report of improvements made.
Potential Rubric(s)	Rubric for application of ML concepts for image recognition – Use stage (Gresse von Wangenheim *et al.*, 2021).
Characteristic features	The task must have the students to clean and label a dataset of recycling trash images. The task must have the students to train the ML model. The tasks must have the students to analyze and interpret the performance of the model based on validation results (accuracy measure and confusion matrix) and testing of new images.
Variable features	No variable features are identified at the Use stage.

interdisciplinary application, the course "ML for All!" guides the students to build an ML model for image classification addressing the topic of recycling thrash. After the motivation and presentation of basic ML concepts and neural networks, the students start the development of the pre-defined ML model (Fig. 1). To build the ML model students are guided step-by-step to use a visual environment, Google Teachable Machine (GTM) (Google, 2023). In addition, students are provided with a set of 210 resized and uncategorized images in order to prepare a dataset. Students have to clean the dataset and label the images with respect to the recycling categories: metal, paper, plastic, and glass. They are also encouraged to expand the dataset by collecting images of trash they have on hand. Then, they are instructed to train the model with GTM, test the model with new images, and interpret the performance achieved by the model, taking into account their tests, the accuracy of the model, and the confusion matrix provided by GTM. During the course, the students are also instigated to



a) Dataset preparation

b) Model training and adjustments

```
{
“type”:”image”,
“version”:”2.4.4”,
“appdata”:
  {“publishResults”:
      {“tensorflowjs”:
          {“name”:”aa_Omited”,
          “token”:”bb_Omited”
          }
      },
  “trainEpochs”:50,
  “trainBatchSize”:16,
  “trainLearningRate”:0.001
  }
}
```

c) Training parameters assessed directly from Google Teachable Machine file (.tm)

Vamos avaliar a acurácia do seu modelo

1. Qual a acurácia do seu modelo?

Ex: 0.03 | Ajuda para encontrar?

2. Qual a acurácia de cada classe?

Ajuda para encontrar?

Metal	
Papel	
Plástico	
Vidro	

3. Analisando os valores de acurácia, você pode observar que

4. Isto indica o que?

d) Example of an online report for evidence collection about interpretation and analysis of accuracy per category

Fig. 1. Examples of work products created by students as a result of learning.

adjust the dataset and/or change the training parameters in order to improve the performance of the ML model. The course is available online for free in Brazilian Portuguese at `https://cursos.computacaonaescola.ufsc.br/`.

c) **Evidence model.** Given a performance in the form of the student’s work products from the tasks, the evidence model details how the information about the student model variables should be updated (Mislevy *et al.*, 2003). Elements of potential observation are elicited in conformity with the student model (Table 3).

The evidence model includes an evaluation model and a measurement model. The evaluation model describes how to extract observable variables in terms of students’ performance from work products of specific tasks and form evidence reflecting students’ information competency level. Observable variables describe characteristics to be evaluated and possibly can group other observable variables together (Mislevy *et al.*, 2003). Thus, in this article, the terms observable variables, items, and rubric criteria will be used interchangeably. Focusing on performance-based assessment the evaluation model is represented in form of a scoring rubric (Table 4) initially proposed by Gresse

Table 3

Evidence model for the assessment of ML competencies at Use stage

Element	Description
Potential observations	Size, distribution, and correctness of the labels of the dataset Execution of the model training Correctness of the performance analysis and interpretation (accuracy table, confusion matrix) Execution of improvement actions Correctness of the analysis and interpretation of prediction tests

von Wangenheim *et al.* (2021), which was reviewed and adjusted, defining the observable variables to be measured to assess the ability to develop an ML model indirectly inferring the achievement of ML competencies. The observable variables in the previous and initial version of the rubric were evaluated in terms of validity by a group of experts and showed a substantial inter-rater agreement as well as content validity in terms of correctness, relevance, completeness, and clarity (Gresse von Wangenheim *et al.*, 2021). With the aim to automate the assessment process, only items that can be assessed automatically were considered. Performance levels were defined according to learning outcomes, specifying the criteria associated with learning objectives, and indicators describing each level to assess student achievement. Higher levels represent greater understanding with respect to the concept being measured. The performance levels were defined on a 4 or 3-point ordinal scale, ranging from *not submitted* to *good* in adherence with the performance expected to achieve the respective learning goal. In case of a lack of submission of certain work products by the student the lowest (not submitted) performance level is associated, assuming that the student has not executed the respective task and, therefore, has not achieved any performance regarding this criteria.

Table 4

Rubric for the assessment of the application of ML concepts for image recognition – Use stage

ID	Item / Observable variables	Performance levels			
		Not submitted – 0 points	Poor – 1 point	Acceptable – 2 points	Good – 3 points
Data management					
I1	Quantity of images	No GTM file (.tm) submitted for assessment	Less than 20 images per category	21 to 35 images per category	More than 35 images per category
I2	Distribution of the dataset	No GTM file (.tm) submitted for assessment.	The number of images in each category varies greatly. More than 10% variation in at least one category (relative to the total)	The number of images between the categories varies between 3% and 10%	All categories have the same amount of images (less than 3% variation)

Continued on next page

Table 4 – continued from previous page

ID	Item / Observable variables	Performance levels			
		Not submitted – 0 points	Poor – 1 point	Acceptable – 2 points	Good – 3 points
I3	Labeling of the images (Sampling 10% of images to test through hi-accuracy ML model)	No GTM file (.tm) submitted for assessment.	Less than 20% of the images were labeled correctly	20% and 95% of the images were labeled correctly	More than 95% of the images were labeled correctly
Model training					
I4	Training	No GTM file (.tm) submitted for assessment.	The model was not trained	The model was trained using the default parameters	The model was trained with adjus-ted parame-ters (epochs, batch size, learning rate)
Interpretation of performance					
I5	Analysis of accuracy per category	No information submitted about categories and/or interpretation.	Categories with low accuracy were not identified	--	All categories with low accuracy were correctly identified
I6	Interpretation of the accuracy	No information submitted about categories and/or interpretation.	Incorrect interpretation of the accuracy analysis of the model	--	Correct interpretation of the accuracy analysis of the model
I7	Analysis of the confusion matrix	No information submitted about Confusion Matrix and/or interpretation.	Incorrect identification of classification errors (more than 2 errors)	Incorrect identification of one or two classification errors	Correct identification of all classification errors
I8	Interpretation of the confusion matrix	No information submitted about Confusion Matrix and/or interpretation.	Incorrect interpretation of the confusion matrix analysis of the model	--	Correct interpretation of the confusion matrix analysis of the model
I9	Adjustments / Improvements made	No information submitted about improvements.	No new development iterations were reported	A new iteration with changes in the dataset and/or training parameters was reported	Several iterations with changes in the dataset and/or training parameters have been reported
I10	Tests with new objects	No information submitted about Tests and/or interpretation.	No new object tested	1–3 objects tested	More than 3 objects tested
I11	Analysis of test results	No information submitted about Tests and/or interpretation.	Incorrect indication of the number of errors in the tests	--	Correct indication of the amount of errors in the tests
I12	Interpretation of test results	No information submitted about tests and/or interpretation.	Wrong interpretation of test results	--	Correct interpretation of test results

The measurement model calculates an overall score ranging from 0.0 to 10.0 in accordance with the grading system typically adopted in Brazil (Santos *et al.*, 2018), as the average of the sum of points of the rubric multiplicated by ten.

Table 5

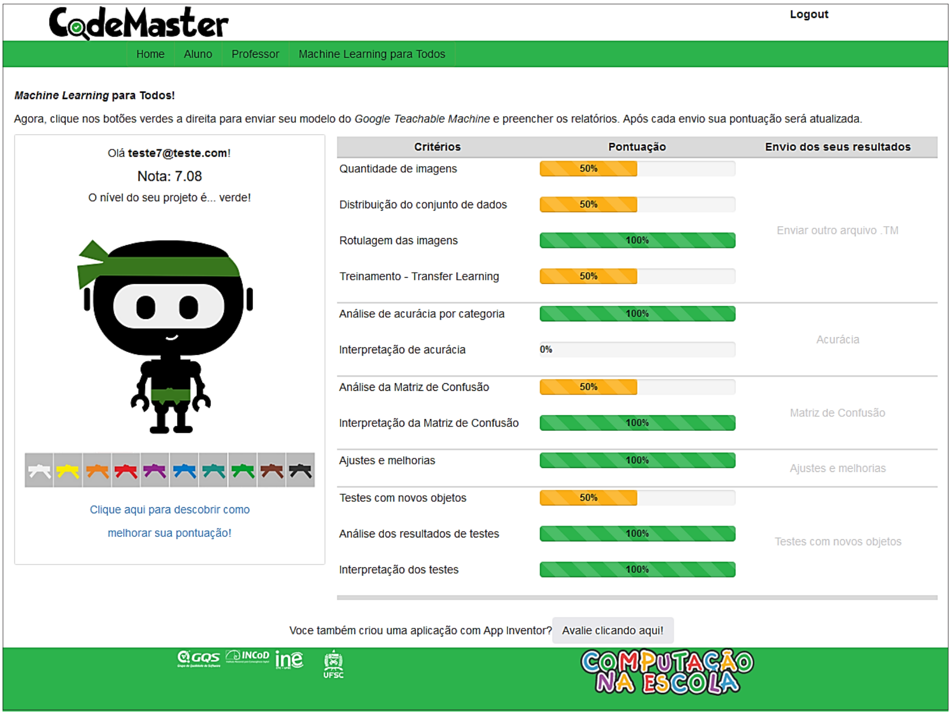
Overview on the main constructs of the Student, Evidence, and Task model

Student model	Evidence model		Task model
Understanding of basic concepts about neural networks. / Ability to collect, clean and label data for the training of an ML model.	I1 I2 I3	Quantity of images Distribution of the dataset Labeling of the images	Dataset preparation
Ability to train an ML model for image classification using a visual tool.	I4	Training	Training
Ability to analyze and interpret the performance of the trained ML model and to improve the model.	I5 I6 I7 I8 I9	Analysis of accuracy per category Interpretation of the accuracy Analysis of the confusion matrix Interpretation of the confusion matrix Adjustments / Improvements made	Performance evaluation
Ability to test the ML model with new images for prediction.	I10 I11 I12	Tests with new objects Analysis of test results Interpretation of test results	Prediction

An overview of the relation between the main constructs of the Student model identifying focal knowledge and skills, the Evidence model (evaluation model) identifying items and the Task model identifying the proposed learning activities is presented in Table 5.

Assessment implementation and delivery. During the application of the course "ML for All!" (Gresse von Wangenheim *et al.*, 2020), work products created by the students were collected as learning outcomes and assessed using the evidence model. Students were instructed to document the results of the ML process during and after the tasks, which includes, besides submitting the generated model by the GTM (.tm) file, reports completed online that document the analysis and interpretation of the performance and prediction results. All these work products resulting from the process of developing the ML model were collected as a basis for the performance-based assessment of the student's learning.

a) **Automation of the assessment – CodeMaster tool.** The assessment model has been automated as part of CodeMaster (Gresse von Wangenheim *et al.*, 2018), a web-based tool for automatically assessing App Inventor apps (and BYOB/Snap! projects) (Fig. 2). The tool was evolved to also automate the assessment of ML concepts based on the evidence model, through the analysis of the work products created as a result of the educational tasks during the course. The work products are submitted iteratively online by uploading the respective artifacts and reports. All submitted data is analyzed and assessed instantly, providing immediate feedback to the students. In order to reduce the processing time for analyzing the labeling correctness with a high-precision deep learning model, we assess only a 10% sample of the images used by the student for training for the assessment of the item "I03 Labeling of the images". The overall score is shown as a numerical value as well as in a



***Machine Learning* para Todos!**

Agora, clique nos botões verdes a direita para enviar seu modelo do *Google Teachable Machine* e preencher os relatórios. Após cada envio sua pontuação será atualizada.

Olá **teste7@teste.com**!

Nota: 7.08

O nível do seu projeto é... verde!

Clique aqui para descobrir como melhorar sua pontuação!

Critérios	Pontuação	Envio dos seus resultados
Quantidade de imagens	50%	Enviar outro arquivo .TM
Distribuição do conjunto de dados	50%	
Rotulagem das imagens	100%	
Treinamento - Transfer Learning	50%	
Análise de acurácia por categoria	100%	Acurácia
Interpretação de acurácia	0%	
Análise da Matriz de Confusão	50%	Matriz de Confusão
Interpretação da Matriz de Confusão	100%	
Ajustes e melhorias	100%	Ajustes e melhorias
Testes com novos objetos	50%	Testes com novos objetos
Análise dos resultados de testes	100%	
Interpretação dos testes	100%	

Voce também criou uma aplicação com App Inventor? Avalie clicando aqui!

Fig. 2. CodeMaster – ML use stage assessment.

ludic form of a robot ninja, whose belt color varies accordingly. In addition, the criteria of the evidence model are listed with the respective performance level achieved. The tool is available in Brazilian Portuguese online for free at `http://apps.computacaonaescola.ufsc.br/codemaster/`.

3. Research Methodology

The research was conducted in an exploratory manner based on a series of case studies, based on data collected from applying the “ML for All!” (Gresse von Wangenheim *et al.*, 2020) course in practice. Following the Goal Question Metric approach (Basili *et al.*, 1994), the objective of this study was to evaluate the assessment model in terms of reliability and construct validity for the performance-based assessment of ML competencies related to image recognition from the researchers’ perspective in the educational context in middle and high school. Based on this objective, the following analysis questions focusing on the scoring rubric as part of the assessment model are derived:

Q1. Is there evidence of the quality of the rubric in terms of difficulty, discrimination, and differentiation?

Reliability

Q2: Is there evidence of internal consistency in the rubric?

Construct Validity
Q3: Is there evidence of convergent and discriminant validity in the rubric?
Q4. How do underlying factors influence the responses on the items of the rubric?

This research was approved by the Ethics Committee of the Federal University of Santa Catarina (No. 4.893.560 and No. 5.610.912).

3.1. *Data Collection*

The sample is composed of learning outcomes collected from middle and high students enrolled in the course, using a non-probabilistic sampling in each case study applying the convenience sampling method (Trochim and Donnelly, 2008). We collected data from seven applications of the course "ML for All!" from 2021 to 2022 (Table 6). The course was applied as an extracurricular activity in six cases and one at a public school as part of school classes. A total of 240 students submitted, sometimes only partially, the work products created throughout the course. Due to the COVID-19 pandemic, most applications were run remotely via Google Meet by an instructor with ML expertise. Two applications were taught face-to-face, with the students present in the school's computer lab. In those applications, the instructor with ML experience taught the content remotely via Google Meet, and the school class teacher acted as an assistant. Basic concepts were taught through interactive lectures, while the practical activities were executed individually by the students following step-by-step instructions available as online material with the assistance of the instructors. One application was carried out asynchronously online by students without an instructor.

Table 6
Overview of the course applications and demographic distribution

Application	Date	Application site/ organizing institution	Instruction mode	Instruction type	Age (years)	Educational stage	No. of students	Middle School (≤15y)	High School (>15y)	Female	Male
AP1	Sep 2021	Public School Dilma Lúcia dos Santos	Face-to-face (instructor remote)	As part of school classes	15-16	Middle school	12	9	3	3	9
AP2	Oct 2021	Federal Institute Catarinense (IFC) Campus Camboriú	Remote instructor-paced	Extracurricular	15-17	Middle and High school	10	1	9	6	4
AP3	Nov 2021	Open to any interested student organized by the Federal University of (UFSC)	Remote instructor-paced	Extracurricular	12-18	Middle and High school	35	9	26	9	26

Continued on next page

Table 6 – continued from previous page

Application	Date	Application site/ organizing institution	Instruction mode	Instruction type	Age (years)	Educational stage	No. of students	Middle School (≤15y)	High School (>15y)	Female	Male
AP4	Mar 2022	Open to any interested student organized by the Federal University of (UFSC)	Remote instructor-paced	Extracurricular	14-18	Middle and High school	38	6	32	15	23
AP5	Sep 2022 to Nov 2022	Non-profit educational organization for underprivileged students Father Vilson Groh Institute	Face-to-face (instructor remote)	Extracurricular	13-18	Middle and High school	109	38	71	47*	59*
AP6	Nov 2022	Open to any interested student organized by the Federal University of (UFSC)	Remote instructor-paced	Extracurricular	15-24	High school and undergraduate	25	2	23	10	15
AP7	Sep 2021 to Dec 2022	Online course	Remote self-paced	Extracurricular	≤18	Middle and High school	11	6	5	**	**
		Total					240	71	169	90†	136†

Note. * 3 students preferred not to indicate their gender at AP6. **Information on gender was not collected as part of AP7. †Considering AP1–AP6.

Once the data was collected, the work products were automatically assessed following the evidence model. In the applications AP5 and AP6, the work products were collected, stored and instantly assessed with the CodeMaster tool, in an automated and anonymous way. For all other applications, the same work products were collected and stored in an online form and later analyzed using the exact same algorithms.

3.2. *Data Analysis*

All the collected data were compiled into a single sample for analysis. Grouping the data was possible due to the similarity of the case studies and the standardization of the data collected by the assessment model, as the case studies were similar in terms of definition, research design, and context. In addition, all case studies were standardized in terms of measures, data collection method, and response format. Typical data preparation for statistical analysis was conducted (Bennett and von Davier, 2017; Rust *et al.*, 2020), grouping categories with low variability. Observing that some students did not submit all the work products, we decided to disregard the data from students who could not be assessed with regard to at least 4 of the items of the assessment model following the procedure proposed by Raghunathan (2004).

Analysis of the quality of items. In order to analyze the quality of the items of the assessment model, we used classical test theory (also called Item Analysis). Item Analysis procedures refer to a set of statistical measures used to review and revise items, estimate their characteristics, and make judgments about the quality of items, typically involving measures of difficulty, discrimination, and differentiation (Bichi, 2016; Bennett and von Davier, 2017; Rust *et al.*, 2020). In order to proceed to the analysis of the difficulty of the items, we use the difficulty index, which is calculated by the proportion of correct answers on each item. In order to proceed to the analysis of the discrimination, the discrimination index is calculated as the difference between the proportion of correct answers of the participants with the higher ability (27% of respondents with higher scores) from those with the lower ability (27% of respondents with lower scores). The biserial correlation is a measure of item differentiation that measures the correlation of the score of a particular test item with the test score. To conduct difficulty and discrimination Item Analysis, due to the nature of the data with items with polytomous responses, the items were considered dichotomized, with the correct answer corresponding to the highest performance level (Acceptable or Good) of each item.

Reliability and validity analysis. In order to evaluate the reliability and construct validity of the assessment model, we used different statistical methods. As reliability refers to the degree of consistency or stability of the assessment items on the same quality factor, we estimate internal consistency by calculating Cronbach's alpha and Omega coefficient (DeVellis, 2017). Both are used to estimate reliability but are determined in different ways. The coefficient omega uses in its calculations a matrix of item loadings on the single common factor that the items share, while Cronbach's alpha derives variance estimates from the covariance (or correlation) matrix of the items (DeVellis, 2017). Even though the most commonly used is the alpha coefficient, the coefficient Omega makes the calculations more stable, with a higher level of reliability and independent of the number of items in the instrument (Flora, 2020), and, thus, both will be considered together here.

Construct validity, on the other hand, refers to the ability that the assessment items manage to measure the latent trait that it proposes to measure, involving convergent and discriminant validity (Brown, 2015; DeVellis, 2017). Convergent validity is the collection of evidence of similarity between measures of theoretically related constructs, while discriminant validity is the absence of evidence of similarity between measures of unrelated constructs (DeVellis, 2017). For both, the degree of correlation between the instrument's items was calculated using the polychoric correlation matrix, which is best suited to ordinal categorical items (Lordelo *et al.*, 2018; Mukaka, 2012). Complementarily, convergent validity was also evaluated by analyzing the correlation between each item and all others, through the item-total correlation (Henrysson, 1963).

We also performed factor analyses (Brown, 2015; DeVellis, 2017) to analyze construct validity, obtaining evidence of convergent and discriminant validity through indicators of factor loadings associated with underlying factors. Following the analysis proposed by Brown (2015), we first checked the suitability of the data for factor analysis through the Kaiser-Meyer-Olkin test. Next, we used parallel analysis, whose approach is based on a scree plot of the eigenvalues obtained from the sample data against eigenvalues that are estimated from a data set of random numbers. Then we performed an exploratory factor analysis in order to find out the essential structure of multivariate observation variables and to identify how much each item is correlated to each sub-dimension through its factor loadings. Due to the nature of the data (items with polytomous responses), the Graded Response Model (Samejima, 1969, 1997; Paek and Cole, 2020) was used for the exploratory factor analysis. For running the data analyses, we used the R language (R Core Team, 2022).

4. Data Preparation

Analyzing the performance achieved based on the work products created by the students throughout the "ML for All!" course and applying the assessment model (Table 4), the frequencies in the performance levels achieved by the students are summarized in Table 7.

We proceed with typical data preparation for statistical analysis (Bennett and von Davier, 2017; Rust *et al.*, 2020). In this regard, items "I05 Accuracy analysis per category", "I06 Interpretation of the accuracy", "I08 Interpretation of the Confusion Matrix", "I11 Analysis of test results", and "I12 Test Interpretation" were recoded to belong to categories with lower codes and sequentially starting at zero points, as the standard for conducting statistical tests. Similarly, item "I03 Labeling of the images" at performance

Table 7
Frequency distribution of achieved performance levels per rubric item

Items		Performance Levels				Total
		Not submitted 0 points	Poor 1 point	Acceptable 2 points	Good 3 points	
I01	Quantity of images	30	77	58	75	240
I02	Distribution of the dataset	30	16	102	92	240
I03	Labeling of the images	30	1	25	184	240
I04	Training	30	0	195	15	240
I05	Analysis of accuracy per category	40	16	-	184	240
I06	Interpretation of the accuracy	40	21	-	179	240
I07	Analysis of the confusion matrix	52	71	41	76	240
I08	Interpretation of the confusion matrix	52	21	-	167	240
I09	Adjustments / Improvements made	39	38	94	69	240
I10	Tests with new objects	39	11	68	122	240
I11	Analysis of test results	39	77	-	124	240
I12	Interpretation of test results	39	47	-	154	240

Note. – Performance level not existent according to the rubric (Table 4).

Table 8

Frequency distribution of performance levels by rubric items after adjustments

Items		Performance Levels				Total
		Not submitted 0 points	Poor 1 point	Acceptable 2 points	Good 3 points	
I01	Quantity of images	17	77	58	75	227
I02	Distribution of the dataset	17	16	102	92	227
I03	Labeling of the images	17	26	184		227
I04	Training	17	195	15		227
I05	Analysis of accuracy per category	29	16	182		227
I06	Interpretation of the accuracy	29	21	177		227
I07	Analysis of the confusion matrix	39	71	41	76	227
I08	Interpretation of the confusion matrix	39	21	167		227
I09	Adjustments / Improvements made	30	38	92	67	227
I10	Tests with new objects	32	11	68	116	227
I11	Analysis of test results	32	74	121		227
I12	Interpretation of test results	32	47	148		227

Note. Not all items have four performance levels, according to the rubric (Table 4).

level "poor" presents and maintains low variability after adjustments, and, thus, was grouped with the level Acceptable. Given the large number of students who did not submit all the work products that are part of the tasks and the potential inaccuracy to be inserted when keeping their data (Raghunathan, 2004), we decided to disregard all data from students whose learning could not be analyzed with regard to at least four of the assessment items. Table 8 shows the results of these adjustments considering the submissions of 227 students.

5. Results

5.1. *Is there Evidence of the Quality of the Rubric in Terms of Difficulty, Discrimination, and Differentiation?*

Initially, analyzing the overall score, the performance achieved by the students was concerning the work products submitted that were assessed according to the items of the evidence model. Fig. 3 presents the percentage of students performing by item correctness. The performance achieved considering all the items dichotomized by their highest category (Acceptable or Good). The average performance achieved regarding all 12 items was 6.2 with a standard deviation of 2.5 and a median of 6. This indicates that on average most of the students achieved a good overall score, an indication that the rubric used in the evidence model is appropriate for the task model.

The difficulty of each item of the evidence model (Table 9) was analyzed considering the difficulty index, given in percentages. As the assessment model uses ordinal

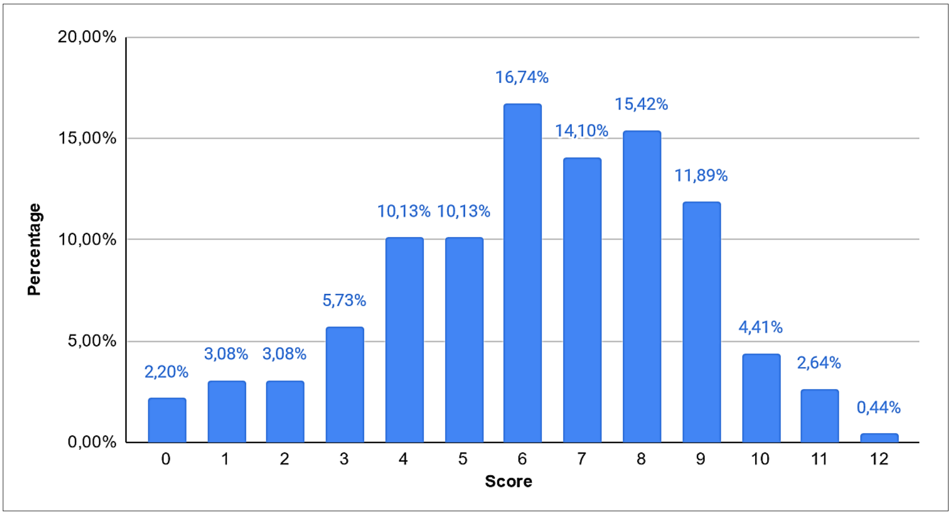


Fig. 3. Students' performance.

polytomous items, these were considered dichotomized in their highest category, where a value higher or close to 100% indicates that the item is easier, while the opposite indicates that the item is more difficult (Bennett and von Davier, 2017; Rust *et al.*, 2020). There is no clear consensus about range limits classification, but ideally, the difficult index for each item should not lie below 25% or above 75%, averaging 50% for the entire items collection (Rust *et al.*, 2020). The difficulty index is typically classified into ranges, so we defined an index from 0 to < 15% very hard, from 15 to < 35% hard, from 35% to < 65% medium, 65% to < 85% easy, and 85% or more as very easy. In general, the items have an adequate difficulty index, with an overall average difficulty index of 52.13%. Items "I03 Labeling of the images" and "I05 Accuracy analysis per category" are slightly easier, but still acceptable, while item "I04 Training" actually turned out to be very hard, demonstrating that few students changed any of the default training parameters (e.g., epochs, batch size, learning rate) of the ML model submitted for assessment.

Additionally, the discrimination index was calculated to estimate the discrimination ability of the items (Table 9). This index ranges from -1 to 1 and is calculated as the difference between the percentages of responses given by the top group and the bottom group (Bennett and von Davier, 2017; Rust *et al.*, 2020). There is no clear consensus on the classification of the range limits, but items with negative values indicate that deletion or revision of the item should be considered (Bichi, 2016; Bennett and von Davier, 2017; Rust *et al.*, 2020). The discrimination index rating is assumed as excellent for 0.3 or above, adequate between 0.2 and < 0.3, moderate from 0.1 to < 0.2, low from 0 to < 0.1, and not discriminating, if less than zero. The results of our analysis show that all items are adequate with respect to discrimination, with the majority being excellent, which consistently indicates that a higher proportion of the group of students with the highest overall performance obtained the highest level of performance for the items.

Table 9

Item quality according to classical test theory

Item		Difficulty index	Classification	Biserial	Classification	Discrimination index	Classification
I01	**Quantity of images**	33.0	Hard	**0,005**	**Inadequate**	**0,029**	**Low**
I02	Distribution of the dataset	40.5	Medium	0,267	Adequate	0,242	Adequate
I03	Labeling of the images	81.1	Easy	0,498	Excellent	0,341	Excellent
I04	**Training**	6.6	Very Hard	0,223	Adequate	**0,051**	**Low**
I05	Analysis of accuracy per category	80.2	Easy	0,85	Excellent	0,485	Excellent
I06	Interpretation of the accuracy	78.0	Easy	0,808	Excellent	0,496	Excellent
I07	Analysis of the confusion matrix	33.5	Hard	0,488	Excellent	0,423	Excellent
I08	Interpretation of the confusion matrix	73.6	Easy	0,821	Excellent	0,561	Excellent
I09	Adjustments / Improvements made	29.5	Hard	0,42	Excellent	0,361	Excellent
I10	Tests with new objects	51.1	Medium	0,311	Excellent	0,321	Excellent
I11	Analysis of test results	53.3	Medium	0,495	Excellent	0,466	Excellent
I12	Interpretation of test results	65.2	Easy	0,492	Excellent	0,444	Excellent

And, although only items “I01 Quantity of images” e “I04 Training” demonstrated a low discrimination level, they can still be considered, as only items with a discrimination below zero should be excluded.

In order to measure differentiation, which is also related to discrimination, we calculate the biserial correlation (Table 9). Its main advantage over the discrimination index is that by considering all responses and not just groups at the edges, this test has a greater power of discrimination (Bichi, 2016). The biserial correlation value is in the range between -1 and 1. It is expected that the highest performance levels (Acceptable or Good) will have the highest and positive Biserial Correlation values, and gradually, for the lower levels, it should decrease until the lowest level (Not submitted), for which it is expected to be negative (Bennett and von Davier, 2017; Rust *et al.*, 2020). There is no clear consensus on the classification of the range limits, but items with negative values indicate that deletion or revision of the item should be considered (Bichi, 2016; Bennett and von Davier, 2017; Rust *et al.*, 2020). The Biserial Correlation rating for the highest performance level is assumed as excellent for 0.3 or above, adequate between 0.2 and < 0.3, moderate from 0.1 to < 0.2, inadequate from 0 to < 0.1, and inappropriate, if less than zero. Most of the items of the assessment model were classified regarding differentiation as Excellent and two as Adequate. Furthermore, all “Not Submitted” item levels show a negative biserial correlation, as expected. This implies that the assessment is adequate since students with high total scores (considering all items) are the students assessed on “Good” or “Acceptable” performance levels on these items, and, at the same time, implies that students with low total scores are the students assessed on low-performance levels. Only item “I01 Quantity of images” demonstrated a low differentiation but still considered appropriate, since its performance level “Good” is close to zero.

5.2. *Is there Evidence of Internal Consistency in the Rubric?*

Reliability was analyzed with respect to the internal consistency of the rubric, calculating Omega and Cronbach's alpha coefficients. Both are used to estimate internal consistency, which indicates how well an instrument's items are correlated. There is no clear consensus on the classification of the range limits, but typically Omega and Cronbach's alpha coefficients above 0.70 indicate the internal consistency of the instrument (values between 0.7 and 0.8 are acceptable, values between 0.8 to 0.9 indicate good, and values greater than or equal to 0.9 indicate excellent internal consistency) (Brown, 2015; Bennett and von Davier, 2017; Rust *et al.*, 2020). The analysis of the rubric resulted in a coefficient Omega of 0.834 and Cronbach's alpha of 0.83. Both coefficients point in the same direction, indicating a good internal consistency of the rubric.

Analyzing whether the internal consistency increases when removing an item (Table 10), it can be observed that the Omega coefficient increases slightly when eliminating some items ("I01 Quantity of images'', "I02 Distribution of the dataset", and "I04 Training"). The coefficient Cronbach's alpha increases slightly when eliminating the item "I01 Quantity of images". The results could point out the exclusion of these items. But this is not supported for several reasons: both coefficients are already at adequate levels, considering the inclusion of the items in question, and, although excluding some items may lead to an increase this increase is negligible, thus their exclusion is not sensibly significant. In addition, the exclusion of items in question would negatively affect other properties of the rubric, especially the measure of difficulty and differentiation. As a consequence, the average performance score would increase significantly, since these items are precisely the most difficult ones, being classified as very hard, hard, or medium. Thus, their exclusion would eventually lead to a very easy instrument, focusing on assessing only artifacts from lower-performing students, which would flatten the upper part of the performance in the graph (Fig. 3.). Another reason is that considering

Table 10
Omega and Cronbach's alpha coefficients when removing an item

Item		Omega when removing an item	Cronbach's alpha when removing an item
I01	Quantity of images	**0.85**	**0.84**
I02	Distribution of the dataset	**0.84**	0.82
I03	Labeling of the images	0.82	0.82
I04	Training	**0.83**	0.83
I05	Analysis of accuracy per category	0.78	0.81
I06	Interpretation of the accuracy	0.78	0.81
I07	Analysis of the confusion matrix	0.81	0.81
I08	Interpretation of the confusion matrix	0.79	0.80
I09	Adjustments / Improvements made	0.82	0.82
I10	Tests with new objects	0.83	0.82
I11	Analysis of test results	0.82	0.81
I12	Interpretation of test results	0.82	0.82

the student model as well as the learning objectives, it is important that students understand that the accuracy of the model improves significantly in function of the number of images used for training, allocated to each category and adjusting the training parameters. This illustrates the importance of maintaining all these items.

5.3. *Is there Evidence of Convergent and Discriminant Validity in the Rubric?*

Convergent and discriminant validity were analyzed by means of the degree of correlation between the items of the instrument based on their polychoric correlation matrix (Fig. 4). It is expected that the items that are measuring a single dimension present correlations greater than or equal to 0.30 (DeVellis, 2017). In this regard, correlations (r) whose value in modulus does not exceed 0.5 ($0.30 \leq | r | < 0.50$) are considered a weak linear correlation, up to 0.7 ($0.50 \leq | r | < 0.70$) represent a moderate correlation, and above ($0.70 \leq | r | < 0.90$) a strong or ($| r | \geq 0.90$) very strong correlation (Mukaka, 2012).

In Fig. 4, positive correlations are shown with the background highlighted in blue, in a gradient that tends toward dark blue as the degree of correlation increases. It can be observed that there are several pairs of items that show a correlation above 0.3 as expected, many of them strong and moderate, which indicates a statistical relationship in the association between pairs of items. The strongest correlation with a value of 0.95 was demonstrated between items I05xI06 which refer to the correct analysis and interpretation of the accuracy of the model categories performed by the student. Following, with 0.85, is the correlation of the pair I08xI07, which refers to the student's analysis and interpretation of the confusion matrix of the ML model. Regarding convergent validity, at least two clusters have been identified, the first one referring to the "Data management" and "Model training" dimension of the rubric, encompassing item I01

		I01	I02	I03	I04	I05	I06	I07	I08	I09	I10	I11	I12
Quantity of images	I01	1.00											
Distribution of the dataset	I02	0.56	1.00										
Labeling of the images	I03	0.49	0.81	1.00									
Training	I04	0.49	0.59	0.79	1.00								
Analysis of accuracy per category	I05	0.03	0.26	0.44	0.16	1.00							
Interpretation of the accuracy	I06	0.04	0.28	0.36	0.18	0.95	1.00						
Analysis of the confusion matrix	I07	0.04	0.24	0.29	0.19	0.71	0.68	1.00					
Interpretation of the confusion matrix	I08	0.07	0.36	0.35	0.19	0.82	0.84	0.85	1.00				
Adjustments / Improvements made	I09	0.05	0.07	0.09	0.15	0.51	0.51	0.49	0.64	1.00			
Tests with new objects	I10	-0.11	-0.03	-0.06	-0.14	0.49	0.47	0.48	0.37	0.47	1.00		
Analysis of test results	I11	-0.13	0.06	0.15	0.14	0.54	0.51	0.49	0.52	0.55	0.76	1.00	
Interpretation of test results	I12	-0.14	0.04	0.07	-0.03	0.50	0.53	0.46	0.54	0.48	0.77	0.81	1.00

Fig. 4. Polychoric correlation matrix.

to I04. The second more visible one encompasses items I05 to I12 that are related to "Interpretation of performance".

On the other hand, correlations outside these two dimensions are mostly weak. There are also some negative correlations, although really weak, which indicate that there is an inversely proportional relationship between the pair, not to be expected. In any case, these correlations are insignificant, and the tendency in a larger sample is for them to be adjusted.

Yet, in general, the rubric presents adequate discriminant and convergent validity and confirms an alignment with the pre-established theoretical dimensions, grouping "Data management" and "Model training" into one dimension and items related to "Interpretation of performance" into another dimension.

To complement the previous analysis, we evaluated the correlation between each criterion and all others using the corrected item-total correlation method (Henrysson, 1963) (Table 11). Each item of the instrument should have a medium or high correlation with all other items (DeVellis, 2017), as this indicates that the items are consistently compared to the other items, considering a correlation as satisfactory if the correlation coefficient is greater than 0.29 (Cohen, 1960). Results show that most items of the rubric have a correlation above or near 0.29. Items I01 to I04 which refer to the "Data management" and "Model training" dimension, in general, have a lower item-total correlation, and only item "I01 Quantity of images" has a low correlation, which indicates that the item may not satisfactorily correlated with the other items in the rubric and possibly should be revised. However, as part of the specific domain it seems to be important to be maintained, as the exclusion of the item would negatively affect other properties of the rubric, especially the measure of difficulty and differentiation. And considering the student model as well as the learning objectives, it is important that students realize that the accuracy of the developed model improves significantly in function of the number of images used for training the model.

Table 11

Item-total correlation of the rubric

Item		Item-total correlation
I01	Quantity of images	**0.12**
I02	Distribution of the dataset	0.35
I03	Labeling of the images	0.36
I04	Training	0.30
I05	Analysis of accuracy per category	0.65
I06	Interpretation of the accuracy	0.65
I07	Analysis of the confusion matrix	0.59
I08	Interpretation of the confusion matrix	0.70
I09	Adjustments / Improvements made	0.51
I10	Tests with new objects	0.50
I11	Analysis of test results	0.57
I12	Interpretation of test results	0.54

5.4. *How do Underlying Factors Influence the Responses on the Items of the Rubric?*

In order to obtain evidence of underlying factors that influence the items of the assessment model we performed a factor analysis (DeVellis, 2017). To check the possibility of performing a factor analysis we used the Kaiser-Meyer-Olkin index (Brown, 2015). It measures the sampling adequacy with values between 0 and 1. A value near 1.0 supports a factor analysis and a value less than 0.5 indicates that the data is not likely suitable for a useful factor analysis (Brown, 2015). Analyzing the items of the rubric, we obtained a KMO index of 0.78, demonstrating that factor analysis is suitable in this case.

Next, we used parallel analysis, a method for determining the number of components or factors to retain which factors with eigenvalues greater than 1 may be significant (DeVellis, 2017). Fig. 5 shows the results of the scree plot with two eigenvalues above the red line. This suggests the existence of two underlying factors or traits in the sample.

We proceeded to exploratory factor analysis (Table 12), in order to compare the indicators of factor loadings associated with underlying factors. Given the nature of the data, the Graded Response Model was used (Samejima, 1969, 1997; Paek and Cole, 2020). In order to decide which items are loaded in each factor, we use the Oblimin rotation method, in which the factors are allowed to be correlated (Jackson, 2005). The models were trained without any priors, by modifying only the number of dimensions.

When considering a single latent trait, we can observe in Table 12/column “One Dimension” that the factor loadings associated with items “I01 Quantity of images”, “I02 Distribution of the dataset”, and “I04 Training” are below the expected value of 0.30, but items “I02 Distribution of the dataset” and “I04 Training” are very close to it. In this case, only item “I01 Quantity of images” demonstrates a very low factor loading. This suggests that there is more than one latent factor, indicating that one factor was not sufficient to explain the variations among the items, and hence there was a violation of the assumption of unidimensionality.

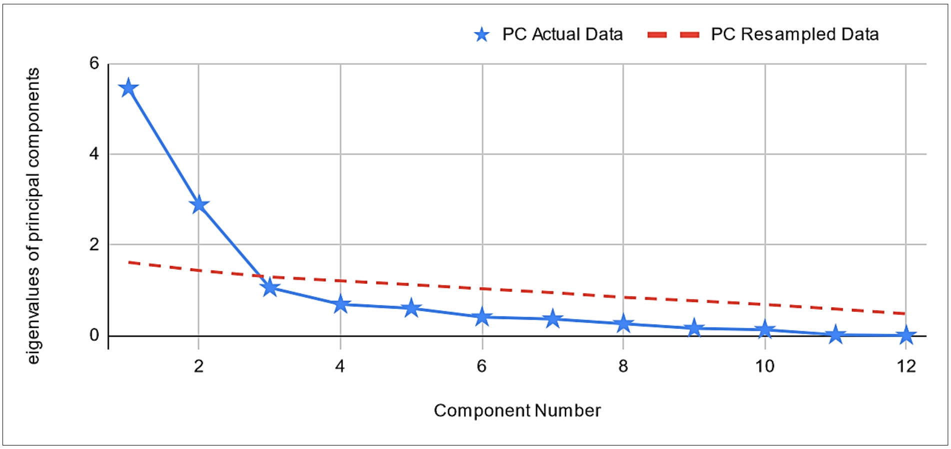


Fig. 5. Parallel analysis scree plot.

Table 12

Exploratory factor analysis of the rubric

Item		One Dimension	Two Dimensions	
		F1	F1	F2
I01	Quantity of images	**0.04**	-0.14	**0.68**
I02	Distribution of the dataset	**0.27**	0.03	**0.91**
I03	Labeling of the images	0.34	0.07	**0.90**
I04	Training	**0.27**	0.02	**0.79**
I05	Analysis of accuracy per category	0.89	**0.81**	0.18
I06	Interpretation of the accuracy	0.89	**0.81**	0.17
I07	Analysis of the confusion matrix	0.84	**0.77**	0.11
I08	Interpretation of the confusion matrix	0.92	**0.83**	0.19
I09	Adjustments / Improvements made	0.63	**0.66**	-0.04
I10	Tests with new objects	0.61	**0.76**	-0.28
I11	Analysis of test results	0.68	**0.80**	-0.16
I12	Interpretation of test results	0.65	**0.78**	-0.21

Note. F1 and F2 denote model dimensions.

When considering two dimensions, in other words, two latent traits, the results in general seem to be more adequate and result in higher values than when considering only one dimension. All items can be included in a dimension with factor loadings well above 0.30, which clearly further confirms the existence of two factors that are being measured (items marked in bold), indicating the same groupings as by the polychoric correlation matrix (Fig. 4). Even item “I01 Quantity of images” presents a good factor loading. Between these two factors, there is a weak correlation of 0.19.

Again, we can observe the alignment with the pre-established theoretical dimensions proposed for the assessment model, grouping “Data management” and “Model training” items into one dimension and another dimension related to “Interpretation of performance”.

5.5. *Threats to Validity*

In order to minimize validity impacts in this study, we identified potential threats and applied mitigation strategies. In order to mitigate threats related to the study design and analysis definition, we defined and documented a systematic methodology following the GQM approach (Basili *et al.*, 1994) and Evidence-Centered Design (Mislevy *et al.*, 2003).

Another issue concerns the quality of the data grouped in a single sample. This was made possible by the standardization of the data, all collected as part of the applications of the “ML for All!” course.

Overcoming the challenges inherent in the automation of the assessment model we adopted Evidence-Centered Design (Mislevy *et al.*, 2003) to systematically define the observable variables that are considered in the automation. In addition, the observable variables in a previous and initial version of the rubric were evaluated in terms of validity by a group of experts and showed a substantial inter-rater agreement as well as content validity in terms of correctness, relevance, completeness, and clarity (Gresse von Wangenheim *et al.*, 2021).

Another risk relates to the validity of the scores inferred based on the work products collected. As our study is limited to assessments using the ML Use rubric, this risk is minimized, as the analyses and inference of the performance levels of the rubric items were performed automatically (using a Python script) with the CodeMaster tool. And, although during the first four applications (AP1–AP4) the CodeMaster tool had not yet been implemented, the same work products were collected through an online form and later analyzed using the exact same algorithms currently implemented in the CodeMaster tool. Exceptions are the items I07 and I08 that were manually analyzed (for applications AP1–AP4) by the authors due to the artifacts being collected as images. For these two items, the evaluation was done manually by one researcher, following exactly the same algorithm used by CodeMaster, and reviewed by a second researcher to reduce the risk of scoring errors.

Another risk is related to the clustering of data from multiple contexts. However, since the goal is to analyze rubric validity in a context-independent manner, this is not considered a problem here, as the course objectives, content and assessment model remained the same across all applications. Another threat to external validity is associated with the sample size and the diversity of the data used. Our analysis is based on a sample of 227 students, and polytomous item analyses for reliability and validity yielded robust results.

Another issue concerns the possible influence exerted by researchers on the data and analysis. In order to mitigate this threat, we adopted a systematic methodology, clearly defining the purpose of the study, data collection, and statistical analysis. Statistical methods were selected with care following the procedure proposed by DeVellis (2017) for the construction of measurement scales, and with procedures typically used for the analysis of internal consistency and construct validity of measurement instruments (Bennett and von Davier, 2017; Rust *et al.*, 2020).

6. Discussion and Conclusion

The main goal of this research was to evaluate the assessment model in terms of reliability and construct validity for the performance-based assessment of ML competencies related to image recognition from the researchers' perspective in the educational context in middle and high school applying the "ML for All!" course in practice. A previous and initial version of the rubric was evaluated by a group of experts and showed a substantial inter-rater agreement of content validity, adequate in terms of correctness, relevance, completeness, and clarity (Gresse von Wangenheim *et al.*, 2021).

The results of the evaluation indicate that the rubric items for assessing ML learning performance shows a good variability of the degree of difficulty and the greatest power of discrimination and differentiation. Thus, demonstrating a great capability to differentiate between better and worse student performance, an important characteristic in any assessment.

In addition, the results indicate that the rubric achieved good levels of internal consistency with a global coefficient Omega of 0.834 and a Cronbach's alpha of 0.83, demonstrating its reliability. Considering the current lack of statistically evaluated performance-based assessment models in this emerging knowledge, the comparison of the evaluation results is very limited. Hitron *et al.* (2019) analyzed the reliability of the coding performed by researchers when manually labeling student's short answer items of an essay on basic ML understanding reporting an interrater Kappa of 92%. Shamir and Levin (2021) report the analysis of content validity, conducted by students and instructors, reviewing the questions for the ability to read and understand the items, but without providing statistical results. Although the objective of the evaluations differs, Hsu *et al.* (2022) analyzed the reliability of a five-item self-assessment questionnaire reporting a Cronbach's alpha of 0.883, which is a "good" range of reliability. These results reported in literature, and considering the theoretical basis inherent from item analysis, demonstrates that reliability of the assessment model proposed here achieves the same levels as the few similar studies.

The polychoric correlation matrix showed moderate to strong correlations for many items evidencing convergent validity. With regard to discriminant validity, the results point to the existence of two dimensions, one, involving item I1–I4, which refers to the dimensions "Data management" and "Model training" and another one including item I5–I12, which refers to the dimension "Interpretation of performance" of the rubric.

Using the proposed evaluation model, we conducted an analysis of underlying factors that influence the items in the evaluation model, seeking to validate the three underlying factors: "Data Management", "Model Training" and "Performance Interpretation". The results, corroborated by the parallel analysis plot and the exploratory factor analysis, clearly show the existence of two underlying factors. We can observe that the factor loadings are adequate and have higher indices when we consider two underlying factors, in which all items can be grouped into one of the two dimensions with good factor loadings, forming the same groups as shown through the polychoric correlation matrix (Fig. 4): one referring to the unification of "Data management" and "Model training", encompassing item I01 to I04 and the second underlying factor encompassing item I05 to I12 related to "Performance interpretation".

All analyses point out item "I01 Quantity of images" to be the most complicated, as well as item I02 and I04 yet a little less severe. However, a possible exclusion of these criteria would have negative effects on other properties of the items. Both coefficients (Omega and Cronbach's alpha) are already at adequate levels considering the inclusion of the items in question, and although exclusion may lead to an increase in the coefficients, this increase would be insignificant. While on the other hand, considering the measures of difficulty and differentiation, an exclusion of these items would lead to a significantly higher mean performance score, since these items are the most difficult. Thus, their exclusion would lead to a very easy instrument, focusing on the assessment of aspects that are commonly achieved in students' work products. Another reason to maintain these items is related to the learning objectives. As pointed out by Gresse von Wangenheim *et al.* (2021) and Martins *et al.* (2023), it is important that students realize that the accuracy of the developed model improves significantly in the function of the number of images used for training, allocated to each category and adjusting the training parameters and, thus, the importance of maintaining these criteria. Perhaps, the pre-trained neural network used in the GTM tool (Google, 2023) that leads to good results even with few images ends up overshadowing this importance to students.

Overall, the results of this evaluation show acceptable reliability and validity of the assessment model to be used for the assessment of building ML models for image classification as part of computing education on middle and high school level applying the "ML for All!" course in practice. In this regard the assessment model can represent an important step in the inclusion of teaching ML in K-12. Based on our experience throughout all applications we observed that the model has the potential to assist in an assessment process. Students themselves can self-assess their work products throughout the "ML for All!" course using the CodeMaster tool to obtain instantaneous feedback and to guide their learning process. It may also reduce teachers' workload on assessment and leave them free to spend more time on other activities with the students, as well as to review the automated assessment as well as to conduct further complementary assessments on factors that are not easily automated.

Automated assessment of ML learning has advantages, such as efficiency and scalability, but also faces challenges such as the complexity of the subject, attempts at manipulation or faking where students may intentionally adjust their work products to appear to have learned more than they actually did, and the lack of consideration of important dimensions. Consequently it is important to complete the automated assessment with the analysis of human instructores in order to provide more robust and meaningful results and to prevent any of these manipulations. Alternatives to complete the automated assessment include interviews, peer reviews, presentations, etc., as suggested for example also in the context of the assessment of the learning of computational thinking (ie., Tang *et al.*, 2020; Kong *et al.*, 2022; Su and Yang, 2023).

Furthermore it is important to emphasize that this assessment model has been systematically developed with regard to the learning objectives and content of the course "ML for All!", and is therefore limited to this context. However, as the learning objectives have been specified in alignment with prominent AI curricular guidelines, the evaluated assessment model may also provide a reliable and valid basis for the development of assessment models in similar course contexts that are also aligned with these curricula.

As part of future work we aim at developing a scale based on further statistical analysis as well as extending the assessment model covering other levels of the Use-Modify-Create cycle.

Acknowledgments

We would like to thank all the students who participated in the course as well as everybody who contributed in some way to their execution.

This work was supported by the CNPq (National Council for Scientific and Technological Development), a Brazilian government entity focused on scientific and technological development.

References

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., Zimmermann, T. (2019). Software Engineering for Machine Learning: A Case Study. *Proc. of IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice,* Montreal, Canada, 291–300.

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

Bichi, A.A. (2016). Classical Test Theory: An Introduction to Linear Modeling Approach to Test and Item Analysis. *International Journal for Social Studies*, 2(9), 27–33.

Basili, V.R., Caldiera G., Rombach H.D. (1994). Goal Question Metric Paradigm. In *Encyclopedia of Software Engineering*, Wiley.

Brown, T.A. (2015). Confirmatory factor analysis for applied research, Second edition. The Guilford Press.

Camada, M.Y. Durães, G.M. (2020). Ensino da Inteligência Artificial na Educação Básica: um novo horizonte para as pesquisas brasileiras. *Proc. of XXXI Brazilian Symposium on Informatics in Education.* Porto Alegre, Brazil, 1553–1562.

Caruso, A. L. M., Cavalheiro, S. A. da C. (2021). Integração entre Pensamento Computacional e Inteligência Artificial: uma Revisão Sistemática de Literatura. *Prof. of XXXII Brazilian Symposium on Informatics in Education*, Porto Alegre, Brazil, 1051–1062.

Cohen, J. (1960). Statistical Power Analysis for the Behavioral Sciences. New York, NY: Routledge.

DeVellis, R.F. (2017). *Scale development: theory and applications*, 4th ed. SAGE.

Flora, D.B. (2020). Your Coefficient Alpha Is Probably Wrong, but Which Coefficient Omega Is Right? A Tutorial on Using R to Obtain Better Reliability Estimates. *Advances in Methods and Practices in Psychological Science*, 3(4), 484–501.

Google. (2023). Google Teachable Machine. Retrieved 09/02/2022 from `https://teachablemachine.withgoogle.com/`

Gresse von Wangenheim, C., Hauck, J.C.R., Demetrio, M.F., Pelle, R., Cruz Alves, N. da, Barbosa, H., Azevedo, L.F. (2018). CodeMaster – Automatic Assessment and Grading of App Inventor and Snap! Programs. *Informatics in Education*, 17(1), 117–150.

Gresse von Wangenheim C., Alves N. da C., Rauber M.F., Hauck J.C.R., 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), 479–500.

Gresse von Wangenheim, C., Marques, L.S., Hauck, J.C.R. (2020). Machine Learning for All – Introducing Machine Learning in K-12, SocArXiv, 1–10. `https://doi.org/10.31235/osf.io/wj5ne`

Hattie, J., Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112.

Hitron, T., Orlev, Y., Wald, I., Shamir, A., Erel, H., Zuckerman, O. (2019). Can Children Understand Machine Learning Concepts?: The Effect of Uncovering Black Boxes, *Proc. of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland, Uk, 1–11.

Henrysson, S. (1963). Correction of item-total correlations in item analysis. *Psychometrika*, 28(2), 211–218.

Ho, J.W., Scadding, M. (2019). Classroom Activities for Teaching Artificial Intelligence to Primary School Students. *Proc. of the Int. Conference on Computational Thinking*, 157–159.

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

Hsu, T.-C., Abelson, H., Van Brummelen J. (2022). The Effects of Applying Experiential Learning into the Conversational AI Learning Platform on Secondary School Students. *The International Review of Research in Open and Distributed Learning*, 23(1), 82–103.

Huba, M. E., Freed, J. E. (2000). *Learner-centered assessment on college campuses: Shifting the focus from teaching to learning*. Allyn & Bacon.

Jackson, J.E. (2005). Oblimin rotation. *Encyclopedia of Biostatistics*. John Wiley & Sons Hoboken, USA. 10(0470011815).

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

Kong, S.-C., Lai, M. (2022). Validating a computational thinking concepts test for primary education using item response theory: An analysis of students' responses. Computers & Education, 187, 104562.

LeCun, Y., Bengio, Y., Hinton G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

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

Long, D., Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proc. of the Conference on Human Factors in Computing Systems*, Honolulu, USA, 1–16.

Lordelo, L.M.K., Hongyu, K., Borja, P.C., Porsani, M.J. (2018). Análise Fatorial por Meio da Matriz de Correlação de Pearson e Policórica no Campo das Cisternas. *E&S Engineering and Science*, 7(1), 58–70.

Lwakatare, L.E., Raj, A., Bosch, J., Olsson, H.H., Crnkovic, I. (2019). A taxonomy of software engineering challenges for machine learning systems: An empirical investigation. *Proc. of the Int. Conference on Agile Software Development*, Montréal, Canada, 227–243.

Marques, L.S., Gresse von Wangenheim, C., Hauck, J.C. (2020). Teaching Machine Learning in school: A systematic mapping of the state of the art. *Informatics in Education*, *19*(2), 283–321.

Martins, R.M., von Wangenheim, C.G., Rauber, M.F., Hauck, J.C. (2023). Machine Learning for All! – Introducing Machine Learning in Middle and High School. *International Journal of Artificial Intelligence in Education*. Online.

MEC. (2018). *Base Nacional Comum Curricular*. Ministry of Education.. Retrieved 01/06/2022 from `http://basenacionalcomum.mec.gov.br/`

MEC. (2020). *Census of Basic Education 2020*. Ministry of Education. Brasília. `https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/notas_esta tisticas_censo_escolar_2020.pdf`

Mislevy, R.J., Almond, R.G., Lukas, J.F. (2003). A Brief Introduction to Evidence-Centered Design. *ETS Research Report Series*, 2003(1), i–29.

Mislevy, R.J., Haertel, G.D. (2006). Implications of evidence-centered design for educational testing. *Educational measurement: issues and practice*, 25(4), 6–20.

Mitchell, T.M. (1997). *Machine Learning*, New York: McGraw-Hill.

Mukaka, M.M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical journal*, 24(3), 69–71.

Paek, I., Cole, K. (2020). *Using R for Item Response Theory Model Applications*. 1 ed. 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/`

Raghunathan, T.E. (2004). What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data. *Annual Review of Public Health*, 25(1), 99–117.

Ramos, G., Meek, C., Simard, P., Suh, J., Ghorashi, S. (2020). Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction*, 35(5–6), 413–451.

Rauber, M.F., Gresse von Wangenheim, C. (2022). Assessing the Learning of Machine Learning in K-12: A Ten-Year Systematic Mapping. *Informatics in Education*, 22(2), 295–328.

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

Rust, J., Kosinski, M., & Stillwell, D. (2020). *Modern Psychometrics: The Science of Psychological Assessment*. 4th ed. Routledge.

Samejima, F. (1969). Samejima, F. (1969). Estimation of latent ability using a response of graded scores. Monograph 17. *Psychometrika*, 34(2), 1–97.

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

Santos P.S., Araujo L.G.J., Bittencourt R.A. (2018). A mapping study of computational thinking and programming in brazilian k-12 education, In: *Proc. of Frontiers in Education Conference*, San Jose, USA, 1–8.

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

Shamir, G., Levin, I. (2021). Neural network construction practices in elementary school. *Künstliche Intelligenz*, 35(2), 181–189.

Su, J., Yang, W. (2023). A systematic review of integrating computational thinking in early childhood education. *Computers and Education Open*, 4, 100122.

Tang, X., Yin, Y., Lin, Q., Hadad, R., Zhai, X. (2020). Assessing computational thinking: A systematic review of empirical studies. *Computers & Education*, 148, 103798.

Touretzky, D., Gardner-McCune, C., Martin, F., Seehorn D. (2019). Envisioning AI for K-12: What Should Every Child Know about AI? *Proc. of* the *AAAI Conference on Artificial Intelligence*, Honolulu, USA, 33(01), 9795–9799.

TIC EDUCAÇÃO. (2019) *TIC Educação*. Cetic. São Paulo, Brazil.

Trochim, W.M.K., Donnelly, J.P. (2008). *The research methods knowledge base*, 3rd ed. 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`

United Nations. (2015). *The 17 Goals.* Department of Economic and Social Affairs, Sustainable Development. Retrieved 06/06/2022 from `https://sdgs.un.org/goals`

Villani, M., Oliveira, D.A. (2018). Avaliação Nacional e Internacional no Brasil: Os vínculos entre o PISA e o IDEB. *Educação & Realidade*, 43(4), 1343–1362.

Yasar, O., Veronesi, P., Maliekal, J., Little, L., Vattana, S., Yeter I. (2016) Computational Pedagogy: Fostering a New Method of Teaching. *Proc. of the Annual Conference & Exposition Proceedings*, New Orleans, USA, 26550.

M.F. Rauber is a professor in the Informatics area at the Instituto Federal Catarinense (IFC), Camboriú. Currently, he is a Ph.D. student in the Graduate Program in Computer Science (PPGCC) at the Federal University of Santa Catarina (UFSC) in Florianópolis, Brazil, and a research student at the Computação na Escola initiative (INCoD/INE/UFSC). He received a M.Sc. (2016) in Science and Technology Education from the Federal University of Santa Catarina, a specialization (2005) in Information Systems Administration from UFLA and a BSc. (2004) in Computer Science from UNIVALI. His main research interests are computing education and assessment.

C. Gresse von Wangenheim is a professor at the Department of Informatics and Statistics (INE) of the Federal University of Santa Catarina (UFSC), Florianópolis, Brazil, where she coordinates the Software Quality Group (GQS) focusing on scientific research, development and transfer of software engineering models, methods and tools and software engineering education. She also coordinates the initiative Computing at Schools, which aims at bringing computing education to schools in Brazil. She received the Dipl.- Inform. and Dr. rer. nat. degrees in Computer Science from the Technical University of Kaiserslautern (Germany), and the Dr. Eng. degree in Production Engineering from the Federal University of Santa Catarina

P.A. Barbetta is a professor at the Federal University of Santa Catarina (UFSC), Florianópolis, Brazil. He graduated in Statistics, master´s degree in Applied Mathematics/Statistics and PhD in Production Engineering. He works in application of statistics in educational assessment, especially in the study of factors associated with performance based on hierarchical regression and construction of measures based on uni- and multidimensional item response theory. He is the author of two textbooks in applied statistics.

A. Ferreti Borgatto is a Professor at the Federal University of Santa Catarina (UFSC). His research interest is statistical analysis in educational assessment, particularly in Item Response Theory (IRT) and missing data analysis. He is currently a consultant at INEP. He teaches the postgraduate program in methods and management in assessment and the postgraduate program in Physical Education. He received a Ph.D. degree in Experimental Statistics at the ESALQ/USP (Brazil) and a Ph.D. degree in Statistics at the UNESP (Brazil).

R.M. Martins is a Telecommunications professor at the Instituto Federal de Santa Catarina (IFSC) in São José, Brazil. Currently, he is a Ph.D. student in the Graduate Program in Computer Science (PPGCC) at the Federal University of Santa Catarina (UFSC) in Florianópolis, Brazil, and a research student at the Computação na Escola initiative (INCoD/INE/UFSC). He obtained his MSc. in Telecommunications from the National Institute of Telecommunications (INATEL) in 2014 and completed specializations in Telecommunications systems in 2015 and Systems Engineering in 2018 from ESAB. He also holds a B.Eng. in Telecommunications from UNISUL. His main research interests include computing education and machine learning.

J.C.R. Hauck is a professor at the Department of Informatics and Statistics (INE) of the Federal University of Santa Catarina (UFSC), Florianópolis, Brazil, where he is co-coordinator of the Software Quality Group (GQS) and the initiative Computing at Schools. He holds a Ph.D. in Knowledge Engineering from the Federal University of Santa Catarina, and his main research interests are in computing education and software engineering.