Informatics in Education, 2022, Vol. 21, No. 1, 55–90

DOI: 10.15388/infedu.2022.03

Does Handwriting Impact Learning on Math Tutoring Systems?

Felipe de MORAIS*, Patricia A. JAQUES

Graduate Program in Applied Computing (PPGCA),
Universidade do Vale do Rio dos Sinos (UNISINOS)
São Leopoldo, Rio Grande do Sul, Brazil
e-mail: felipmorais@edu.unisinos.br, pjaques@unisinos.br

Received: February 2021

Abstract. Intelligent Tutoring Systems (ITSs) for Math still use traditional data input methods: computers' keyboard and mouse. However, students usually solve math tasks using paper and pen. Therefore, the gap between the manner the students work and the requirements imposed by these typing-based systems expose students to an extraneous cognitive load, impairing their learning. Our study investigates the impact of the data input method on students' learning and fluency in solving equations using step-based math ITSs. More specifically, we have considered the standard typing and handwriting input methods. We hypothesized that the students would be more fluent using their handwriting with online recognition to solve math equations than using the typing input method. This fluency indicates a reduction in cognitive load, freeing working memory for logical reasoning instead of interface preconditions, leading to improved learning. We have conducted an experiment with 55 seventh-grade students from a private school to validate the hypothesis, randomly assigned to control and experimental groups. Each group used one of the input methods on two different devices (desktop computers and tablets). Although students using handwriting solved more equations and were faster than students who typed their equations, we could not find statistically significant differences in the learning between students that used typing or handwriting. Additionally, we have found that the input method used in a not ideal device (e.g., handwriting with a computer's mouse instead of using a touch screen device) can negatively affect the students' performance.

Keywords: handwriting, intelligent tutoring system, cognitive load.

1. Introduction

Activities with greater interactivity, enabling students to do rather than just visualize, result in improved learning (Koedinger *et al.*, 2015). Intelligent Tutoring Systems (ITS) are educational software that can provide this interactivity, as they can assist students

*Corresponding author.

individually in solving their tasks, helping to improve their learning (Koedinger *et al.*, 1997). Unlike other educational software, ITSs are experts in a particular field, and they can provide fine-grained personalized assistance, being able to adapt the instruction according to the learner's knowledge (Graesser *et al.*, 2016). They have been seen to be almost as effective as individual human tutoring (VanLehn, 2011).

ITSs in the math field require students to enter the solution of the tasks in a specific manner that the computer can understand. One example of these tasks is the first-degree equations. These equations are sentences that contain an equality relation and use a mathematical notation represented by symbols, numbers, and letters. They have only one variable, for which the student should identify a value to balance the equation.

To solve first-degree equations, the students must apply basic mathematical operations (addition, subtraction, multiplication, and division). However, it is not always simple to type mathematical expressions in a computer-based learning environment due to keyboard input restrictions, often requiring students to follow a complex set of steps (Anthony *et al.*, 2012). For example, to insert a fraction into an equation in a particular ITS, the student must follow 13 steps. This 13-step process may seem intuitive for those who have been using the system for a long time, called experienced users, but this is considered a complicated and time-consuming task for novice students.

The cognitive load theory explains that working memory limits human cognitive processing. This memory can handle a limited amount of information, thus "high cognitive load can hamper learning and transfer" (Sweller *et al.*, 2019). Cognitive load is associated with the complexity of the content, or information, to be studied, called intrinsic cognitive load. Certain content related to various other contents is considered highly complex, and it has a high intrinsic cognitive load. Cognitive load is also associated with how the content is presented to the learner, called extraneous cognitive load. Real-world interfaces cause students to experience high extraneous cognitive load, which directly affects the speed at which the task is performed, the focus of attention, meta-cognitive control, accuracy in problem-solving, and memory (Oviatt *et al.*, 2006).

Besides how this content is presented to the learner, the requirements imposed by the instructional procedure also enhance the cognitive load (Sweller, 2010). Thus, the more the student needs to reason about how the content is presented, the less s/he will reason about the content to be learned. In ITSs, the system's graphical user interface can impose extraneous cognitive load, which can impose a task-irrelevant to the student goal of solving an equation. Thus, by exposing the student to complex and lengthy processes to use the interface, the system imposes an additional extraneous cognitive load. It causes the student to occupy her or his working memory with restrictions and procedures of the user interface rather than reasoning about solving the equations themselves (Oviatt *et al.*, 2006).

One possible way to reduce students' extraneous cognitive load using a computer-based learning environment would be to "design interfaces that are more similar to existing work practice" (Oviatt *et al.*, 2006). Since the beginning of school classes, students use pencils and paper to solve mathematical problems to draw and freely represent their thoughts on paper using the handwriting method. As this process is repeated continuously throughout the student's school life, handwriting becomes an automated task, leading the brain to save cognitive resources, which could be used to process higher-level tasks (Wicki *et al.*, 2014). We have used the term **fluency** to define the student's ability to easily and fluently enter the equations in the system, i.e., doing without thinking about it.

According to Read *et al.* (2001), students can write more fluently using handwriting when compared to the typing method. However, most of the math ITSs found in the literature require the students to use traditional computer devices, such as the keyboard and mouse, to insert their equations on the system. The problem with the typing approach is that students experience difficulties using the keyboard because the text-like equation confuses them by requiring a set of unknown patterns. Thus, instead of reasoning about solving the equation, the student has to think about how to use the ITS (Read *et al.*, 2000).

Researches on computer-based learning environments using the handwriting as input method have already found that: *i*) the handwriting input method is faster than typing (Anthony *et al.*, 2005, 2007b), *ii*) students can solve more problems in the same time (Glaser, 1976), compared to other methods, and *iii*) students prefer the handwriting (Lee *et al.*, 2012), leading to increased engagement (Elliott and Dweck, 1988). However, there is still no evidence that the handwriting method can improve learning. Also, based on a literature review, we were unable to find research using a step-based math ITS to provide feedback on every solving step, neither providing real-time handwriting recognition during the use of the system, which has been applied to a performance and fluency evaluation, comparing the handwriting and the typing methods. Another differential of this work is related to the device used for handwriting and typing the equations in the ITS. We have verified whether these methods used with different input devices, e.g., touchscreen or digital pen versus mouse for handwriting, could affect the students' fluency and performance.

This work investigates the impact of the typing and handwriting data input methods on a step-based ITS over students *i*) learning and *ii*) fluency in solving first degree equations. We hypothesized that, by allowing the students to use their handwriting with real-time recognition as the equation input method in a step-based ITS for math, they would solve the equations more naturally, focusing their reasoning on solving the equation instead of the requirements for using the graphical user interface. Therefore, this approach could lead to improved learning. We have also investigated whether the different devices used to input data, i.e., tablet device and computer with mouse and keyboard, could impact the results of each input method. This work extends the results from Morais and Jaques (2017) and Morais *et al.* (2017b) by presenting a whole new experiment and discussions.

We have conducted an experiment to test our hypothesis, in which the students were randomly assigned to control and experimental groups. Students in the experimental group have used the handwriting input method, and the students in the control group have used the typing input method to solve the equations on the ITS PAT2Math. Besides the input method, the experiment was divided into two phases. The students in phase 1 have used the tablet's touchscreen to handwriting (experimental group) and the computer's keyboard to type (control group). In contrast, during phase 2, we exchanged the input device while maintaining the input method: the students in the experimental group used the computer's mouse for handwriting, and the students in the control group typed their equations into the system using the tablet's keyboard. We have chosen this approach to verify the device's impact on the students' performance and fluency.

2. Literature Review

This section aims to present an overview of the learning environments that have been using the handwriting input method. Some papers have been published to present or improve handwriting recognition accuracy (Anthony *et al.*, 2012; Tran Minh Khuong *et al.*, 2019). However, the goal of this work is not related to presenting a novel way or improved manner of handwriting recognition. We seek to investigate the impact of different input methods and devices on students' learning and fluency to solve the equations.

We have searched the ACM, IEEE, Elsevier, and Springer digital libraries to identify learning environments that have been using the handwriting method to replace the typing method. Our search was based on the variations of the keywords handwriting, sketch-based interfaces, tutoring systems, and learning environments. We have restricted the search for papers published in the last five years, i.e., between 2015 and 2020. After finding a set of papers, we have applied one snowballing phase, searching for relevant papers that were not included in the first set. At this time, we had no date restrictions.

We have found computer-based learning environments that use the handwriting or sketch-based methods as the data input to improve the learning of the students for a variety of domains, such as music theory (Barreto *et al.*, 2016), physics (Cheema and LaViola, 2012; Lee *et al.*, 2012; Cheema and LaViola Jr., 2018), Japanese kanji writing (Taele and Hammond, 2009), drawing techniques (Dixon *et al.*, 2010; Cummmings *et al.*, 2012), mathematics (Laviola, 2007; Anthony *et al.*, 2007a; Vuong *et al.*, 2010; Anthony *et al.*, 2012; Pacheco-Venegas *et al.*, 2015; Phon-Amnuaisuk *et al.*, 2015; Wang *et al.*, 2016), biology (Taele *et al.*, 2009), essentials of writing (Thompson *et al.*, 2016), engineering statics concepts (Valentine *et al.*, 2015), freehand engineering sketching (Hilton *et al.*, 2019), digital circuit design (Alvarado *et al.*, 2015), and geometry (Kang *et al.*, 2016, 2017). From this set of works, we have selected only the papers that are more similar to the research goal of our work and the domain of equations solving. Therefore, the following paragraphs describe works that have focused on replacing the keyboard and mouse with the handwriting/sketch-based input method to improve the usefulness of the system in improving the students learning.

WebMath is a web-based tool that enables students to insert math problems in a web browser using their handwriting through a digital pen or the computer's mouse (Vuong *et al.*, 2010). The tool can recognize the handwriting and display it back to the student, which can be edited if any error is found. After recognizing the handwriting, WebMath provides a step-by-step solution to the problem. However, as the tool is not an ITS, it cannot help students solve the problems, providing step-by-step guidance. In this paper, the authors discuss the implementation, architecture of the tool, and an experiment focused on handwriting recognition accuracy.

Cognitive Tutor Algebra I is an ITS that provides step-by-step feedback and helps students solving math problems. Anthony and colleagues integrated a handwriting recognition tool into the tutor's interface for students to enter equations into the system (Anthony *et al.*, 2007a). This tool was based on a device attached to the monitor, allowing the students to use a stylus pen to write on the screen. After handwriting all the solving steps of the equation, the students were requested to type the final answer of the equation. The ITS provided feedback just when the student entered the final answer in the system. The authors reported that students who used the handwriting to enter answers in the tutor were twice faster than students who used the keyboard and mouse (Anthony *et al.*, 2007b). However, no learning gain difference was found between the two groups of students (Anthony *et al.*, 2007b). Requesting the students to type the final answer was a strategy to guarantee the correct feedback from the ITS. The authors chose this strategy because the handwriting recognition accuracy was not ideal. Therefore, in later studies, they have focused on improving recognition accuracy (Anthony *et al.*, 2012).

Newton's Pen II (NP2) (Lee *et al.*, 2012) is an ITS for physics, which uses online handwriting recognition to provide a more natural interface. In the solution part of the physics equations, the student writes the solving steps using a stylus pen, and the system checks the steps entered by the student. An evaluation was conducted with more than 100 engineering students, in which students answered a survey with questions about system usability and its usefulness for learning. The questions obtained above-average answers, illustrating students' acceptance and preference for the proposed system.

MathDIP is a web-based tool developed to help students during math problem-solving (Pacheco-Venegas *et al.*, 2015). Although the system cannot provide specialized error feedback, the tool can assist students step-by-step, providing automatic evaluation and *correct-incorrect* feedback, according to a given set of available solutions. Also, it allows the students to enter equations with the handwriting input method. A qualitative study was conducted in which the students expressed an overall acceptance of the system.

Wang *et al.* (2016) presented Math Tutor, an application to assist students in solving linear and quadratic equations by using the handwriting input method on an Android tablet. Although Math Tutor is not considered an ITS, it can provide specialized messages for wrong expressions to help the students during problem-solving. The authors conducted a qualitative study to identify the user experience and get the students' feedback about the tool. As a result, students enjoyed using the application, and they would use it for assignments if available. The authors also reported that students became more efficient and less prone to error as they used the application.

AnalyticalInk (Kang *et al.*, 2016) is an environment that uses an interactive online handwriting-recognition interface that seeks to help students understand geometric concepts and solve algebra and geometry exercises. The system provides problems in a textual form, with some highlighted keywords. The student uses this information to solve the problems and can drag the information to the solving part of the equations. It can provide visual feedback on students' handwritten steps using a tablet and a stylus pen. A qualitative evaluation with ten students who used the system showed that the system is helpful to guide with the geometric and algebraic problems solving, and the students were satisfied.

All related works have reported the usage of the handwriting modality for entering information into the systems. This work differs from related ones in three main aspects: *i*) it allows the student to see and edit her or his recognized handwriting in real-time, *ii*) it provides specialized feedback at each solving step of the equation, and *iii*) it does not require the use of a particular device. Another significant difference is related to the research goal. Most of the related work performed qualitative research, aiming to verify the student's acceptance over the handwriting modality. This work aims at verifying the impact of the typing and handwriting input methods on students learning and fluency. An additional contribution of this work is on verifying whether this impact is related to the device used to perform the input, i.e., tablets or desktop computers. Also, the experiment goals of the related works and this work are slightly different. This paper aims to verify whether the handwriting input reduces the extraneous cognitive load of students. Most of the related works did not conduct an experiment, controlling the input methods. Besides, they have not considered the effect of the devices on the results.

3. The Intelligent Tutoring System PAT2Math

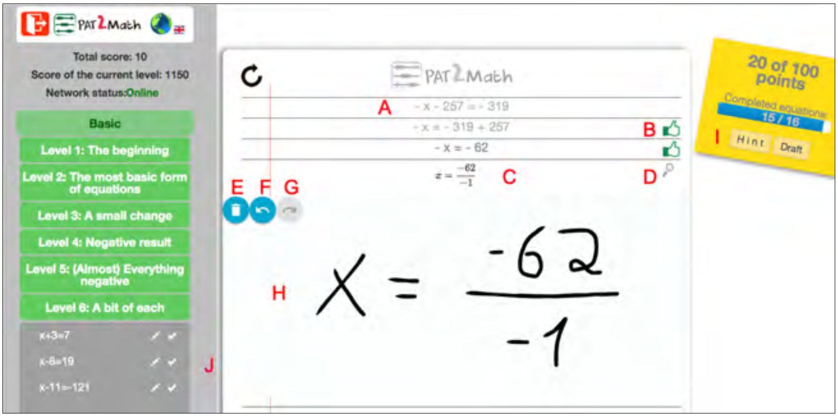
This section describes the ITS used by the students to solve the math tasks, called PAT2-Math. The PAT2Math (Personal Affective Tutor to Math) is an intelligent tutoring system that assists students in the task of solving first-degree equations (Jaques *et al.*, 2013). PAT2Math runs as a server-side application with a web-based graphical user interface, allowing students to use the system on any computer or device through a web browser. Thus, every interaction between the student and the ITS is sent to the server, processed, stored, and returned to the student.

According to Vanlehn's classification (Vanlehn, 2006), PAT2Math is considered a step-based ITS because it contains an inner loop that can provide individualized assistance to the students for every solving step of a given equation, called minimal feedback. Therefore, the ITS provides specialized feedback for each step; if it is correct, the student can enter the next step; if it is wrong, the ITS provides error feedback. Besides the feedback, PAT2Math also provides scaffolding hints whenever the student requests them.

By default, to solve a given equation in the PAT2Math's interface, the student must use the computer keyboard and type the solution step in an input text box. After typing the step, the student can either hit the *enter* key or push the verify button. Once one of these events is fired, the input box gets blurred, the text is converted to a readable equation, and the step is sent to the server for checking. If the step is correct, the system replaces the input box with the two-dimensionally rendered equation, and it inserts a new input box for the next step. This process is repeated until the student provides the final correct answer for the equation. To allow the students to use their handwriting to enter math equations into the system, we have replaced the traditional input box with a handwriting box. We have developed this new handwriting input box as a plugin to the PAT2Math user interface, allowing it to be accessed by smartphones, tablets, or computers via a web browser.

This plugin integrates two works: the PAT2Math ITS and the online handwriting recognizer MyScript *Math*. The PAT2Math ITS provides all the feedback and specialized hints to assist the student during the whole solving process of an equation. On the other side, MyScript provides a web development kit that receives the student's writing and converts it into formatted text (Mouchère *et al.*, 2014). So, when the student writes in the handwriting box, the plugin sends the data to the MyScript server that extracts static and dynamic handwriting information, inserts that information into a combination of a Deep Multi-Layer Perceptron and a Recurrent Neural Network, and gets a list of probability symbols[1]. After a spatial relationship analysis of the writing and this list, MyScript returns the recognized handwriting in a pre-defined text format. Upon receiving this text, the developed plugin interprets the formatted text and displays the equation for the student in a multidimensional manner. MyScript is available based on a commercial license and has been used by thousands of applications worldwide. In this study, we have used the MyScript API based on a free educational account.

Fig. 1 shows a screenshot of the developed handwriting plugin integrated into PAT2-Math's graphical user interface. To use the handwriting input method, the student has to access the tool from a mobile device or a computer, using a web browser, and s/he needs to login to PAT2Math with her or his user account and choose the handwriting as input mode. When accessing a plan, a list of first-degree equations will be displayed (Fig. 1.J). If the student has already solved any of them, it will be saved in the PAT2Math' student module and can be reviewed again. If the student has not solved the equation yet, s/he



Total score: 10
Score of the current level: 1150
Network status:Online

$-x - 257 = -319$
$-x = -319 + 257$
$-x = -62$
$x = \frac{-62}{-1}$

$x = \frac{-62}{-1}$

20 of 100 points
Completed equations 15 / 16

Fig. 1. Screenshot of the handwriting tool integrated into PAT2Math ITS interface.

[1] Winner of the ICFHR 2016 CROHME competition, MyScript has achieved an accuracy of 92,81% for math symbols recognition (Mouchère *et al.*, 2016).

can select the chosen equation, and it will be displayed at the top of the screen as the initial equation (Fig. 1.A). Below the initial equation, there is a handwriting input box where the student can use his or her finger (or a stylus pen) to write the steps to solve the equation (Fig. 1.H). As s/he writes the numbers and mathematical symbols, the tool displays what has been recognized in real-time (Fig. 1.C).

Once the solving step has been written, the student must touch the verify button (Fig. 1.D) to inform the system that the step is ready to check. The tool then sends the written step in the text-like format to the tutor module, which forwards it to the expert module for correction and saves the information into the student module. The feedback is returned to the tutor module that sends it back to the interface. If the answer is correct, the tool provides visual feedback (Fig. 1.B), and it allows the student to proceed to the next solving step and so on until s/he enters the final correct answer of the equation. The student can also request help from the ITS by touching the "Hint" button (Fig. 1.I), which will provide a specialized hint message to the student. If the step is incorrect, the system allows the student to edit the writing in real-time, including some writing help buttons (Fig. 1.E (clean the handwriting input box), 1.F (undo the last writing), 1.G (redo the last action)). Besides, the plugin allows the student to use shortcuts to facilitate the editing of handwriting. For example, the student can cross out a mathematical symbol to erase it.

4. Method

This research aims to verify the impact of the input data modality on students learning and fluency to solve math tasks. We are also interested in identifying whether the device used to insert the data into the system can influence the results. Therefore, we have considered two different input modalities: typing and handwriting. We hypothesized that by allowing the students to use their handwriting to enter the solving steps of a math problem in the system, with step-by-step guidance and real-time handwriting recognition, they would solve the equations more fluently. Being able to be more fluent would result in the students solving more equations in less time, focusing their reasoning on solving the equation instead of the requirements for using the graphical user interface. Thus, the system's extraneous cognitive load would be reduced, freeing the students' working memory, leading to improved learning (Oviatt *et al.*, 2006).

We have also considered two different devices for the students to use: desktop computers, handled through mouse and keyboard, and tablets, handled by touchscreen. We hypothesized that the students on the experimental handwriting condition could be more engaged and excited to solve the tasks because of the tablets than students on the traditional typing condition utilizing the computer's keyboard, interfering with the results. We also want to verify which device is more appropriate for each input method and their impact on the students' fluency and performance.

This section aims to provide details about the experiment we have conducted to validate our hypotheses. Therefore, this section describes the participants, procedure, experimental design, research instruments, and measures of this experiment.

4.1. *Participants*

We have invited 55 students from two seventh-grade classes of a private school from southern Brazil to participate in our study. The students have used the ITS weekly to solve first-degree equations during one regular math class period of 50 minutes. These students were between 12 and 13 years old (average = 12.15, median = 12, and standard deviation = .36), 26 boys and 29 girls. To perform the experiment and collect the data from the students, we have asked their parents or tutor to sign an informed consent form. The ethics committee of our university validated and approved the form, which contains all the information about the tool, describing how it works and how it would affect the students.

Although the students were using the system as a classroom aid tool at the school's initiative, we only used the data from the students who returned the consent form signed by their parents and by themselves. The data from two students, who did not return the form, were excluded from the analysis. From the 53 remaining students, we removed the data from 10 more students who did not perform one of the knowledge tests or who missed more than one day of data collection. Thus, to generate the results, we have used the data from 43 students. Four students who have missed just one class had their data included in the analysis.

4.2. *Procedure*

The experiment was conducted in nine 50-minutes sessions, spread between May to July 2018, occurring once a week. The first day's goal was only to meet the students and explain to them our research goal and how it would proceed. In all meetings, the teacher of the classes was present. The teacher was the same for both classes and supported the whole experiment.

We had 11 iPads available from our university. As the number of students in the experimental group was greater than the number of tablets, we have asked the students to bring their own devices, if possible. Thus, the experimental group students used 15 devices, ranging between brands, versions, sizes, and operational systems.

4.3. *Experimental Design*

The experiment's goal is to measure the performance of the students based on knowledge tests, as well as the fluency of the handwriting compared to typing, based on the time spent and the number of solved tasks. We were also interested in verifying whether the device (tablet or computer) could impact the student's performance and fluency, given the handwriting or typing methods. Therefore, we have designed the experiment so that all the students could use both devices. We randomly assigned the students into two groups, called experimental and control groups. The students in the control group always used the typing input method to enter the equations on the ITS, regardless of the

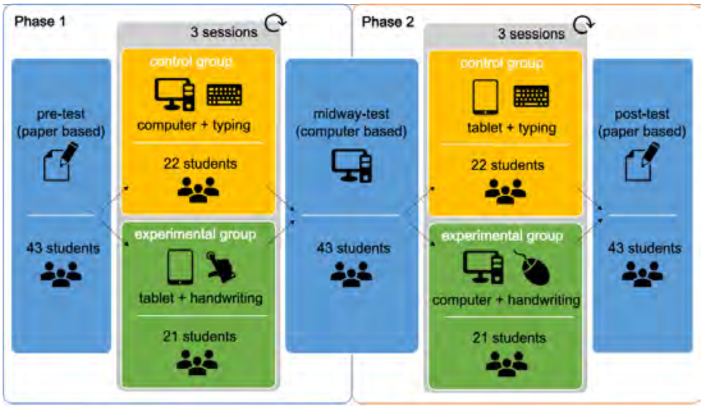


Fig. 2. Experimental design.

device. On the other hand, the students in the experimental group always used the handwriting input method to enter the equations.

As shown in Fig. 2, we have divided the experiment into two phases, called phase 1 and phase 2. The students solved the knowledge tests before, between, and after the phases. The tests do not belong to the phases, but they seek to measure the students' performance in each phase. In phase 1, all 43 students first answered a paper-based test, called pretest. After the pre-test, the students used the PAT2Math ITS during three sessions. The control group, containing 22 students, used the ITS in desktop computers, entering the equations by typing in the keyboard. The experimental group, comprising 21 students, used the ITS in the tablets, entering the equations by handwriting on the devices' screen, i.e., touchscreen with students' fingers (without a stylus pen). After the third session of phase 1, all 43 students answered a computer-based test called midway-test. Based on this pre and midway-tests strategy, it was possible to verify the students' knowledge before and after they have used the ITS in Phase 1. Still, because of the random assignment in control and experimental groups, it was possible to verify whether the students' knowledge was affected by the method of inputting equations into the ITS.

The students performed the midway-test using the ITS. However, there was no feedback from the ITS to the student at this moment. All the students completed the midway-test individually, and they used desktop computers with a keyboard to enter the equations in the ITS. There was no group division during the test; therefore, all of them answered the same test. In this 'test mode', the PAT2Math ITS accepts all the steps from the student without providing any kind of feedback. The pre-test, three sessions of ITS usage, and the midway-test were performed on different days according to the math class schedule.

We have used the midway-test as the upper bound of phase 1 and lower bound of phase 2 for the knowledge variable, which prevented the students from taking two tests in sequence. Thus, after the midway-test, Phase 2 started in which the students used the PAT2Math ITS during three more sessions. However, in Phase 2, although students continued to use the same input method (handwriting or typing), we have changed the input device. The students in the control group used the ITS in the tablet, entering the equations by typing in the device's keyboard. The students in the experimental group used the ITS in the desktop computers, entering the equations by handwriting with the computer's mouse.

After the third session of phase 2, all 43 students answered a paper-based test, the post-test. Based on this strategy of midway-test and post-test, it was possible to verify the students' knowledge before and after they have used the ITS for the second time. Still, due to the random assignment in the control and experimental groups, it was possible to verify whether the input method affected the students' performance on tests, representing their learning. We were also interested in ascertaining the relationship between the input method to enter the equations into the ITS and the devices used to perform this task. Thus, by having the students of the same group using the same input method but with different devices in the different phases, it was possible to verify this relationship between the input method and the device used.

4.4. *Research Instruments*

We have used two types of instruments: the knowledge tests and the PAT2Math ITS. The knowledge tests had the goal of measuring the knowledge of the students in solving first-degree equations. Thus, we have developed the tests to include all the operations needed for the student to master this content. The tests were developed by a partner teacher who has more than 30 years of experience by teaching this content to seventh-grade classes[2]. The tests were developed according to the content the students were learning during the period in which the experiment was performed. In total, there were 14 isomorphic equations[3] for each one of the three tests (pre, midway, and post-tests) for the students to solve. The tests covered questions ranging from the simplest equations (e.g., $x - 8 = -18$) to the most complex equations (e.g., $\frac{x}{3} - \frac{3x}{16} = -10$), given the content of first-degree equations[4].

The complexity of the equations gradually increased in the tests, from the first to the last equation, which allowed the tests to contemplate the need to use multiple operations to solve the equations. Besides, all three tests had the same number of equations with the same complexity distribution to guarantee that the results were measured in the same way for all tests.

The second instrument we have used was the PAT2Math ITS. As the ITS has been described in Section 3, we present here only the equations solved by the students while using the system. We have used the set of original equations already available on the tutor, which the students solved during six sessions. PAT2Math has a gamified graphi-

[2] This teacher is different from the teacher of the classes that participated in the experiment.

[3] We have used the isomorphic term to denote equations that are similar in structure but with different numbers, e.g., $x - 3 = 2$ and $x - 9 = 5$.

[4] The pre, midway, and post-knowledge tests are available in this link:
`https://drive.google.com/file/d/19T8NE3sjfC4IlL94wWQ6br7g1EH7rksI/view?usp=sharing`

cal user interface, in which the contents are organized in phases and levels. Thus, the student has to solve all the equations of the level to unlock the next level. Moreover, the student has to solve all the levels to unlock the next phase. Each level has five equations, except for the review levels, which might have between 10 to 20 equations. The number of levels per phase varies between five to 12, depending on the phase's complexity. The phases and levels are organized to gradually increase the complexity of the equations according to the student's progress within the system. Although the equations in the same level follow the same structure, i.e., they are isomorphic, each level has equations with a more complex structure than its previous levels. Thus, at each level, the student has to solve a more challenging set of equations, requiring the application of a greater number of algebraic operations.

During the experiment, the students solved 200 different equations, starting at the first level of the first phase (e.g., $x + 4 = 9$) and achieving up to level 26 of the third phase (e.g., $9x - (10 - 10x) = 218$). These equations were already available in the PAT2Math ITS from previous experiments. It is essential to highlight that, as PAT2-Math help students individually, it allows students to work at their own pace. Thus,

Table 1

The structure of the equations solved by the students in each phase

	Equation Structure	Example	# Equations
Phase 1	[+−]x[+−]a=[+−]b	x+4=9	70
	[+−](x)/(a)=[+−]b	(x)/(6)=7	31
	[+−]ax=[+−]b	9x=18	30
	[+−]ax[+−]b=[+−]c	3x+10=91	20
	[+−]ax[+−]b=[+−]cx[+−]d	7x+1=6x+6	6
	[+−]ax[+−]b[+−]cx[+−]d=[+−]ex[+−]f[+−]gx[+−]h	14x+20+26x−12=15x−42−5x+8	1
	[+−]ax[+−]b=[+−]x[+−]c	4x+7=x+25	1
	[+−]ax[+−]b=[+−]c[+−]dx	−5x−2=18−3x	1
Phase 2	[+−]x[+−]a=[+−]b	−x−257=−319	37
	[+−]ax[+−]b=[+−]c	3x+10=91	32
	[+−](x)/(a)=[+−]b	(x)/(6)=7	31
	[+−]ax=[+−]b	12x=6	30
	[+−]ax[+−]b=[+−]cx[+−]d	7x+1=6x+6	6
	[+−]ax[+−]b=[+−]x[+−]c	4x+7=x+25	2
	[+−]ax[+−]b=[+−]c[+−]dx	23x−16=14−17x	2
	[+−]ax[+−](b[+−]cx)=[+−]d	9x−(10−10x)=218	2
	[+−]ax[+−]b[+−]cx[+−]d=[+−]e[+−]f[+−]gx	5x−15−4x−8=2+3−6x	1
	[+−]x[+−](ax[+−]b)=[+−]c	x+(2x+5)=35	1
	[+−]ax[+−]b[+−]cx[+−]d=[+−]ex[+−]f[+−]gx[+−]h	14x+20+26x−12=15x−42−5x+8	1
	[+−]x[+−]a[+−]bx[+−]c=[+−]dx[+−]e[+−]fx[+−]g	x+2−3x+4=−5x+6−7x+8	1
	[+−]ax[+−]b[+−]cx=[+−]dx[+−]e[+−]f	10x−5−5x=6x−6−20	1
	[+−]ax[+−]b[+−]c=[+−]d[+−]ex	17x−2+4=10+5x	1
	[+−]ax[+−]b[+−]cx[+−]d=[+−]ex[+−]f	6x−9+2x+2=3x+18	1
	[+−]ax[+−]x[+−]b=[+−]c[+−]x[+−]d	2x−x+1=5−x+3	1

some students stopped way behind level 26. We have grouped all the equations solved by the students in each phase to compare the structure of the equations. This grouping was performed by replacing the numbers in the equations with letters (a, b, c, d,...). This way, we could group equations that have the same structure, i.e., isomorphic equations. Besides, we have considered the different signs, plus or minus, of each number in the same way, using the $[+-]$ symbol. For example, the equations $-4x + 2 = -18$ and $3x - 5 = 7$ have the same structure, only changing the signs. Thus, both equations in this example follow the structure $[+-]\,ax\,[+-]\,b\; = [+-]\,c$. Table 1 shows all the structures computed according to the grouping for phases 1 and 2. It also shows an example of an equation for each structure and the number of equations solved that followed that structure.

As shown in Table 1, we can see that phase 2 has more complex equations than phase 1. However, the number of more difficult equations presented to the student was small compared to more representative structures. In the first four lines of the table, we can see that the vast majority of the equations have the same structure in both phases. In fact, from all the equations solved by the students, only eight in phase 2 followed a different structure from those seen in phase 1. Besides, only one equation from each of these eight structures was solved by the students, representing only 5.3% of the different equations' structures presented to the students during phase 2.

4.5. *Measures*

This section presents the measures we have collected to validate our hypothesis and to calculate the results. In our experimental design, Subsection 4.3, we have used tests to measure students' knowledge before and after the three sessions of each phase. Therefore, the **students' performance** on the tests is our first measure. Each test contained 14 equations worth one point each. The tests were corrected by the same teacher who developed them. Also, all three tests were corrected in the same manner, considering one point if the result and all steps were correct and half-point if some steps were incorrect. For correcting the midway-test (computer-based), we have generated a report from the solution entered into the system by the students. Therefore, the teacher analyzed the students' solution step-by-step, as she did for pre and post-tests. Finally, we have converted the scores of the tests to a 0 to 10 scale, where ten indicates the student correctly solved all the test equations.

By collecting the scores of each student test, we were able to (*i*) calculate the learning gain of each student in each phase after using the ITS, (*ii*) compare the gain between groups in each phase, and (*iii*) compare the gain of the groups between the phases 1 e 2. These calculations would contribute to understanding whether the students learned more after they used the ITS and to identify if students in the experimental condition, who have used the handwriting method, learned more than the students in the control condition, who have used the typing method. Also, it will contribute to verify the impact of the device on the students' learning when comparing the handwriting with the typing input method.

ITS was stored in the database, containing information about the student, the ITS feedback, and a timestamp. This data allowed us to account for all the equations the students have solved and all correct and wrong steps the students have entered in the ITS. This information contributed to the calculation of two other measures: the **number of equations** each student solved and the **number of correct and incorrect steps** in each equation.

The selection of tasks for students to solve while using the PAT2Math is based on a fixed sequence. Therefore, to go to the next level/phase, the student must solve all the previous equations. If s/he got stuck in a given equation/step, s/he could have requested the ITS's help (e.g., by clicking the hint button) or asked the teacher. Thus, there were no unsolved equations. However, the number of solved equations could vary based on each student's knowledge and pace, as described in Section 4.4. In addition to the number of solved equations, we collected the number of correct and incorrect steps that we call just steps for simplification, but we present the results individually. Again, this measure is dependent on the group (control or experimental) and the phase. Therefore, we were able to (*i*) calculate the number of equations/steps solved by each student in each phase while using the ITS, (*ii*) compare the amount of equations/steps solved between groups in each phase, (*iii*) compare the number of equations/steps solved by the groups between phases, and (*iv*) verify the relationship between the number of equations/steps solved with the test scores of the students, for different conditions and phases.

Lastly, we were able to calculate the time spent by each student to enter a step and solve whole equations, called **time on step** and **time on equation**, respectively. We have stored a timestamp for every solution step entered by the student on the ITS. Therefore, we were able to calculate the time difference between the steps entered by the student to solve the equation until the answer to the equation was inserted. This difference represents the time spent for the student to solve the step. The summation of the time spent on steps is equal to the time spent by the student to solve the whole equation. Again, this measure is dependent on the group (control or experimental) and the phase. By computing these time measures, we were able to (*i*) calculate the time spent on each equation/step by each student in each phase while using the ITS, (*ii*) compare the time spent on each equation/step between groups in each phase, (*iii*) compare the time spent on each equation/step by the groups between phases, (*iv*) verify the relationship between the time spent on each equation/step with the number of equations/steps solved, for different conditions and phases, and (*v*) verify the relationship between the time spent on each equation/step with the test scores of the student, for different conditions and phases.

5. Results

This section presents the results obtained from the experiment performed in this study for the measures described in Subsection 4.5. Section 5.1 shows the scores on the performance tests. Section 5.2 presents the findings about the number of solved steps and equations. Furthermore, Section 5.3 describes the results about the time to solve the steps and the whole equation. Moreover, according to the groups and phases, we have added a subsection (Subsection 5.4) to present the correlation between these measures.

To apply any statistical test[5], we have followed the recommendations of Field *et al.* (2012). Thus, we have calculated the Shapiro-Wilk test[6] for every distribution so we could identify, with a p-value $< .05$, which distribution did not follow a normal distribution. Also, for every condition, based on two or more distributions, we have calculated Levene's test for homogeneity of variance[7], so we could identify, with a p-value $< .05$, the distributions in which the variances were significantly different. Any condition involving a non-normal distribution and different variances between distributions violates the parametric assumptions. Thus, for these cases, we have applied a non-parametric test.

By checking for normality and homogeneity of variance of the distributions, we could guarantee the assumptions of parametric tests and, therefore, present more reliable results. This way, when parametric assumptions were not violated, we have applied paired t-tests[8] to check conditions in the same group and independent t-tests to check conditions between different groups. Otherwise, we have used Wilcoxon signed-rank test to check conditions in the same group (i.e., a test to compare means of dependent samples with non-parametric distributions) and the Wilcoxon rank-sum test to check conditions between different groups (i.e., a test to compare means of independent samples with non-parametric distributions). For all statistical analyses, we have considered a significance level of $\alpha = .05$. Also, for each statistical test, we have computed the effect size, which allows verifying whether the effect of the result is important in practical terms. We have followed the equations presented in Field *et al.* (2012) to calculate the effect size for parametric and non-parametric distributions, using Pearson's correlation coefficient r. Also, as suggested by the authors, we have considered $r = .10$ as small effect, $r = .30$ as medium effect, and $r = .50$ as large effect.

5.1. *Tests Performance*

In total, the students have completed three performance tests: pre, midway, and post-tests. Fig. 3[9] shows a box plot with the distribution of the students' scores on each test. Fig. 3 is divided into results obtained from the control and experimental groups, and each group is divided into three tests. The filled dots represent each student's score, the green plus sign indicates the mean of each distribution, the horizontal black bar indicates the median of each distribution, and the red asterisk highlights the outliers of each distribution.

[5] We have imported all results into R Studio (Version 1.1.453) to compute all the calculations and plots.

[6] The Shapiro-Wilk tests were computed using the *stats* R package.

[7] The Levene's tests for homogeneity of variance were computed using the *car* R package.

[8] The t-tests and Wilcoxon tests were computed using the *stats* R package.

[9] The plot and calculations of Fig. 3 were computed using the *ggplot2* R package.

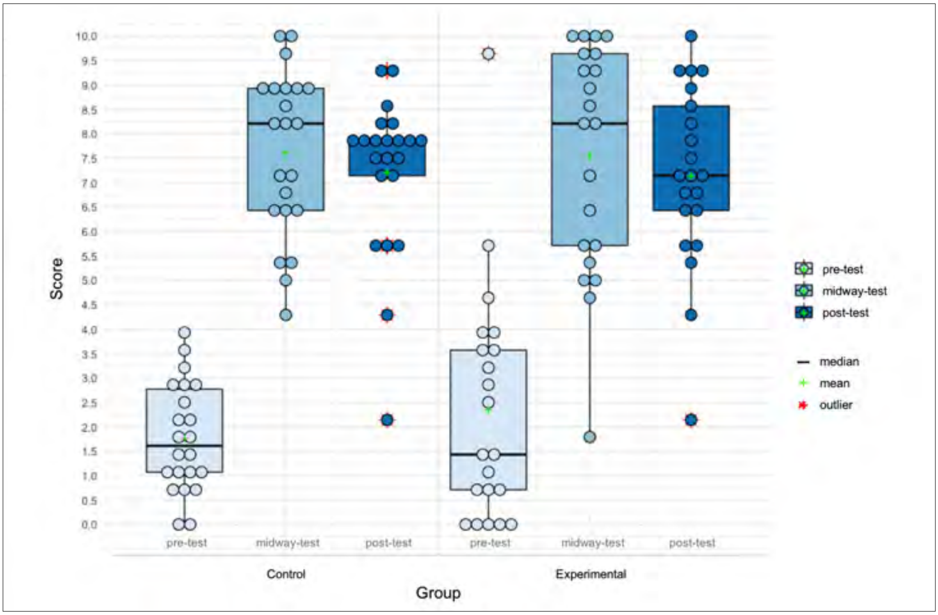


Fig. 3. Distribution of the test scores, according to the group and the test.

Fig. 3 does not highlight the phases, but as stated in Subsection 4.3, the pre and midway-tests are the lower and upper bound of phase 1, respectively, the midway and post-tests are the lower and upper bound of phase 2, respectively, and the pre and post-test are the lower and upper bound of the whole experiment, respectively. Table 2 summarizes the descriptive statistics of the test scores, including, for each test in each group, the number of students, mean, median, standard deviation, standard error, 95% confidence interval, and Shapiro-Wilk normality test of each distribution. We have computed the 95% confidence intervals[10], applying a bootstrap technique, based on 2000 bootstrap replicates, as suggested by Field *et al.* (2012).

Table 2

Descriptive statistics of the test scores

	Test	# Mean	Median	Std.	Dev.	Std. Error	95% conf. int.		Shapiro-Wilk	
							min.	max.	W	p-value
Control	pre	22	1.769	1.61	1.117	.231	1.313	2.229	.959	.470
	midway	22	7.630	8.21	1.683	.356	6.926	8.364	.936	.163
	Post	22	7.224	7.86	1.643	.343	6.547	7.898	.822	**.001**
Experimental	pre	21	2.363	1.43	2.426	.517	1.355	3.387	.857	**.006**
	midway	21	7.550	8.21	2.348	.491	6.608	8.517	.885	**.018**
	post	21	7.143	7.14	1.880	.399	6.350	7.922	.949	.321

[10] The bootstrap confidence intervals were computed using the *boot* R package.

Based on the results of the Shapiro-Wilk tests, presented in Table 2, it is possible to notice that the distribution of the scores of the post-test, $W = .822$, $p = .001$, of the control group and the distribution of the scores of the pre-test, $W = .857$, $p = .006$, and midway-test, $W = .885$, $p = .018$, of the experimental group are significantly non-normal distributions. After calculating the descriptive statistics, we checked the significance of the results for the control and experimental groups and comparison between the groups based on hypothesis tests. These tests are detailed in the following subsections.

5.1.1. *Control Group*

In the control group, we have applied a paired t-test to verify whether the students learned in phase 1, in which the scores on the pre-test (M[11] = 1.769, SE[12] = $.231$) were significantly lower than the scores on the midway-test ($M = 7.630$, $SE = .356$), $t\,(21) = -18.42$, $p < .001$, $r = .97$, with similar variances $F(1, 42) = .12$, ns[13]. For phase 2 of the control group, the results on the midway-test were higher than the results on the post-test.

5.1.2. *Experimental Group*

In the experimental group, we have applied a Wilcoxon signed-rank test to verify whether the students in the experimental group learned in phase 1, in which the scores on the pretest ($Mdn = 1.43$) were significantly lower than the scores on the midway-test ($Mdn = 8.21$), $p < .001$, $r = -.91$, with similar variances $F(1, 40) = .01$, ns. For phase 2, the results on the midway-test were greater than the results on the post-test.

5.1.3. *Between Groups*

After presenting the results about the performance of each group, we have calculated the learning gain of each possible combination of groups and phases. The gain is computed by subtracting the test score performed before the ITS sessions from the score of the test performed after the ITS sessions, which we have called lower and upper bound measurements, respectively. We have computed the mean, median, standard deviation, standard error, 95% confidence interval, based on 2000 bootstrap replicates and Shapiro-Wilk. We have considered two types of learning gain for control and experimental groups: the *midway - pre* represents the learning gain on phase 1, and the *post - midway* represents the learning gain on phase 2. Table 3 presents the descriptive statistics of these results.

After computing the learning gain and the descriptive statistics of the distributions, we have checked for the significance of the results, based on our hypothesis tests. According to the results presented in Table 3, the Shapiro-Wilk normality test showed that all the distributions of the learning gain followed a normal distribution. Thus, we have applied an independent t-test to verify the learning gain difference between the control and experimental groups in phase 1, i.e., *midway - pre*, with similar variances $F(1, 41) = .97$, ns. On average, students in the experimental group achieved lower gains

[11] *M* is used for the abbreviation for Mean.

[12] *SE* is used for the abbreviation for Standard Error.

[13] *ns* is used for the abbreviation for not statistically significant.

Table 3

Descriptive statistics of the learning gain calculated between tests

Gain	Group	Mean	Median	Std. Dev.	Std. Error	95% conf. int.		Shapiro-Wilk	
						min.	max.	W	p-value
midway-pre	Control	5.859	6.07	1.494	.319	5.245	6.454	.939	.188
	experimental	5.187	5.36	2.220	.484	4.248	6.106	.932	.157
post-midway	Control	−.405	−.71	1.875	.399	−1.167	.373	.961	.499
	experimental	−.408	.00	1.635	.357	−1.074	.249	.938	.200

(M = 5.187, SE = .484) than students in the control group (M = 5.859, SE = .319) in phase 1. We have computed the ANCOVA to compare the scores on the midway-test of students in both groups of phase 1. The covariate, the pre-test score, was significantly related to the midway-test $F(1,40)$ = 16.36, p < .05, r = .54. However, there was no significant effect of the group on the midway-test score after controlling for the effect of the pre-test, $F(1,40)$ = .63, p = .43, partial η^2 = .02[14].

We have applied an independent t-test to check the difference in the learning gain between the control and experimental groups in phase 2, i.e., *post - midway*, with similar variances $F(1,41)$ = .49, ns. On average, students in the experimental group achieved lower gains (M = $-.408$, SE = .357) than students in the control group (M = $-.405$, SE = .399) in phase 2. We have computed the ANCOVA to compare the scores on the post-test of students in both groups. The covariate, the midway-test score, was significantly related to the post-test $F(1,40)$ = 20.04, p < .05, r = .58. However, there was no significant effect of the group on the post-test score after controlling for the effect of midway-test, $F(1,40)$ = .01, p = .93, partial η^2 < .01.

5.2. *Number of Equations and Steps*

This section describes the results we have collected for the number of equations/steps entered in the system by the students during the six sessions of using PAT2Math ITS. We also have computed the number of correct and incorrect steps students entered in the system. Table 4 summarizes, for each measure in each group, the number of students (#), mean, median, standard deviation, standard error, 95% confidence interval, based on 2000 bootstrap replicates, and Shapiro-Wilk normality test of each distribution. Based on the results of the Shapiro-Wilk tests, presented in Table 4, it is possible to notice that the number of solved equations, W = .901, p = .031, number of correct steps, W = .895, p = .024, and number of incorrect steps, W = .858, p = .005, only on phase 1 of the control group are significantly non-normal distributions. After the calculation of descriptive statistics, we have checked for the significance of the results, based on our hypothesis tests, for the control group, experimental group, and comparison between groups.

[14] According to Field *et al.* (2012), partial η^2 is an effect size measure for ANCOVA that “looks the proportion of variance that a variable explains that is not explained by other variables in the analysis.”

Table 4

Descriptive statistics on the number of equations and steps performed

Measure	Group	Phase	#	Mean	Median	Std. Dev.	Std. Error	95% conf. int.		Shapiro-Wilk	
								min.	max.	W	p-value
# solved equations	control	phase 1	22	77.14	78.50	15.92	3.394	70.45	83.81	.901	**.031**
		phase 2	22	31.91	34.00	14.64	3.122	25.82	37.82	.938	.180
	experimental	phase 1	21	83.09	82.00	22.10	4.822	73.80	92.38	.971	.755
		phase 2	21	40.09	43.00	14.25	3.109	34.16	46.03	.970	.736
# steps	control	phase 1	22	230.67	239.5	48.43	10.326	211.10	250.00	.953	.358
		phase 2	22	98.00	90.00	36.70	7.824	83.38	112.45	.936	.163
	experimental	phase 1	21	222.67	226.00	68.58	14.964	194.50	250.50	.986	.983
		phase 2	21	105.81	105.00	43.94	9.588	88.30	123.90	.961	.544
# correct steps	control	phase 1	22	184.95	197.00	38.46	8.199	169.60	200.70	.895	**.024**
		phase 2	22	77.82	71.50	35.13	7.490	63.78	92.00	.942	.217
	experimental	phase 1	21	171.76	163.00	58.30	12.722	147.40	196.10	961	.528
		phase 2	21	84.14	74.00	39.40	8.597	67.56	100.57	.945	.273
# incorrect steps	control	phase 1	22	45.68	36.50	25.57	5.452	35.06	56.11	.858	**.005**
		phase 2	22	20.18	17.50	11.77	2.509	15.34	25.00	.962	.539
	experimental	phase 1	21	50.91	49.00	29.03	6.336	38.72	63.14	.916	.071
		phase 2	21	21.67	22.00	9.37	2.045	17.75	25.55	.985	.979

5.2.1. *Control Group*

In the control group, on average, the number of solved equations and the number of entered steps to solve the equations in phase 2 were lower than in phase 1. Thus, based on a Wilcoxon signed-rank test, the number of solved equations in phase 2 ($Mdn = 34.00$) was significantly lower than in phase 1 ($Mdn = 78.50$), $p < .001$, $r = -.641$, with similar variances $F(1, 42) = .065$, ns. Also, based on a Wilcoxon signed-rank test, the number of entered steps in phase 2 ($Mdn = 90.00$) was significantly lower than in phase 1 ($Mdn = 239.50$), $p < .001$, $r = -.640$, with significantly different variances $F(1, 42) = 7.943$, $p = .007$. We have checked the differences in the number of correct and incorrect steps, as well. Thus, based on a Wilcoxon signed-rank test, the number of correct steps in phase 2 ($Mdn = 71.50$) was significantly lower than in phase 1 ($Mdn = 197.00$), $p < .001$, $r = -.641$, with similar variances $F(1, 42) = 3.911$, ns. Also, based on a Wilcoxon signed-rank test, the number of incorrect steps entered in phase 2 ($Mdn = 17.50$) was significantly lower than in phase 1 ($Mdn = 36.50$), $p < .001$, $r = -.596$, with similar variances $F(1, 42) = .187$, ns.

5.2.2. *Experimental Group*

In the experimental group, on average, the number of solved equations and entered steps to solve the equations in phase 2 were lower than in phase 1, as in the control group. Based on a Wilcoxon signed-rank test, the number of solved equations in phase 2 ($Mdn = 43.00$) was significantly lower than in phase 1 ($Mdn = 82.00$), $p < .001$, $r = -.642$, with significantly different variances $F(1, 40) = 6.340$, $p = .016$. Also, based on a Wilcoxon signed-rank test, the number of entered steps in phase 2 (Mdn = 105.00) was significantly lower than in phase 1 (Mdn = 226.50), p < .001, r = $-.642$, with significantly different variances $F(1, 40)$ = 15.009, p < .001. We have checked the differences in the number of correct and incorrect steps. Based on a paired t-test, the number of correct steps in phase 2 (M = 84.14, SE = 8.597) was significantly lower than in phase 1 (M = 171.76, SE = 12.722), $t\,(20)$ = -11.25, p < .001, r = .929, with similar variances $F(1, 40)$ = 1.890, ns. Also, based on a paired t-test, the number of incorrectly entered steps in phase 2 (M = 21.67, SE = 2.045) was significantly lower than in phase 1 (M = 50.91, SE = 6.336), $t(20)$ = -4.67, p < .001, r = .723, with similar variances $F(1, 40)$ = 1.107, ns.

5.2.3. *Between Groups*

After testing the results inside each group, we have tested the hypothesis between groups, for each measure and phase. The number of solved equations of the control group in phase 1 (Mdn = 78.50) was lower than the number of solved equations of the experimental group in phase 1 (Mdn = 82.00). But, a Wilcoxon rank-sum test showed that they did not differ significantly W = 208, p = .292, r = $-.161$, with similar variances $F(1, 41)$ = 2.001, ns. We have checked the same condition on phase 2, in which the number of solved equations of the control group (M = 31.91, SE = 3.122) was also lower than the number of solved equations of the experimental group (M = 40.09, SE = 3.109). An independent t-test showed that this difference in phase 2 was significant $t\,(40.98)$ = -1.858, p = .035, r = .279, with similar variances $F(1, 41)$ = .056, ns.

For the number of entered steps during phase 1, on average, the experimental group (M = 222.67, SE = 14.964) entered fewer steps than the control group (M = 230.67, SE = 10.326). But, based on an independent t-test, this difference was not significant $t\,(35.84)$ = .438, p = .332, r = .073, with similar variances $F(1, 41)$ = 2.692, ns. During phase 2, the control group (M = 98.0, SE = 7.824) entered fewer steps than the experimental group (M = 105.81, SE = 9.588). But, based on an independent t-test, this difference was not significant $t\,(39)$ = $-.631$, p = .270, r = .101, with similar variances $F(1, 41)$ = .533, ns.

We have checked the number of correct steps between groups for phases 1 and 2. During phase 1, the experimental group (Mdn = 163.0) entered less correct steps than the control group (Mdn = 197.0). But, a Wilcoxon rank-sum test showed that they did not differ significantly W = 261.5, p = .233, r = $-.182$, with significantly different variances $F(1, 41)$ = 5.386, p = .025. During phase 2, the control group (M = 77.82, SE = 7.49) entered less correct steps than the experimental group (M = 84.14, SE = 8.597). But, based on an independent t-test, this difference was not significant $t\,(39.96)$ = $-.555$, p = .291, r = .087, with similar variances $F(1, 41)$ = .019, ns.

Finally, we have also checked the number of incorrect steps between groups for phases 1 and 2. During phase 1, the control group (Mdn = 36.5) entered fewer wrong steps than the experimental group (Mdn = 49.0). But, a Wilcoxon rank-sum test showed that they did not differ significantly W = 206.5, p = .280, r = $-.165$, with similar variances $F(1, 41)$ = .166, ns. During phase 2, the control group (M = 20.18, SE = 2.509) entered fewer incorrect steps than the experimental group (M = 21.67, SE = 2.045). But, based

Table 5
Descriptive statistics on time to solve the equations/steps

Measure	Group	Phase	#	Mean	Median	Std. Dev.	Std. Error	95% conf. int.		Shapiro-Wilk	
								min.	max.	W	p-value
time on equation	control	phase 1	22	60.82	62.13	12.50	2.666	55.77	65.89	.953	.355
		phase 2	22	114.60	97.93	41.25	8.794	97.20	131.00	.909	**.045**
	experimental	phase 1	21	57.57	51.81	12.68	2.767	52.31	62.75	.837	**.003**
		phase 2	21	85.39	82.06	25.55	5.575	74.59	96.21	.956	.438
time on a step	control	phase 1	22	20.18	20.35	3.36	.717	18.83	21.54	.979	.905
		phase 2	22	34.71	33.43	8.96	1.910	31.02	38.38	.934	.146
	experimental	phase 1	21	22.11	21.53	5.20	1.134	19.85	24.33	.974	.818
		phase 2	21	32.75	30.98	9.77	2.132	28.68	36.78	.937	.191

on an independent t-test, this difference was not significant $t\,(39.75) = -.459$, p = .324, r = .073, with similar variances $F(1, 41)$ = 1.23, ns.

5.3. *Time to Solve the Equations and Steps*

This section describes the results we have collected about the time that students spent (in seconds) to solve the equations and steps during the six sessions of using PAT2-Math ITS. Table 5 summarizes, for each measure in each group, the number of students, mean, median, standard deviation, standard error, 95% confidence interval, based on 2000 bootstrap replicates, and Shapiro-Wilk normality test of each distribution. Based on the results of the Shapiro-Wilk tests, presented in Table 5, it is possible to notice that the distribution of time spent in equations for the control group of phase 2, W = .909, p = .045, and the experimental group of phase 1, W = .837, p = .003, are significantly non-normal distributions. Again, after the calculation of descriptive statistics, we have checked for the significance of the results, based on our hypothesis tests, for the control group, experimental group, and comparison between groups.

5.3.1. *Control Group*

In the control group, according to a Wilcoxon signed-rank test, the students spent significantly more time solving the equations during phase 2 (Mdn = 97.93) than phase 1 (Mdn = 62.13), p < .001, r = $-$.779, with significantly different variances $F(1, 42)$ = 20.89, p < .001. For the time spent on a step measure, according to a paired t-test, students spent significantly more time on a solving step during phase 2 (M = 34.71, SE = 1.91) than phase 1 (M = 20.18, SE = .717), $t\,(21)$ = 7.842, p < .001, r = .863, with similar variances $F(1, 42)$ = 0, ns.

5.3.2. *Experimental Group*

In the experimental group, according to a Wilcoxon signed-rank test, the students spent significantly more time solving the equations on phase 2 (Mdn = 82.06) than phase 1 (Mdn = 51.81), p < .001, $r = -.649$, with significantly different variances $F(1, 40)$ = 5.917, p = .020. For the time spent to solve a step measure, according to a paired t-test, the students spent significantly more time on a step during phase 2 (M = 32.75, SE = 2.132) than phase 1 (M = 22.11, SE = 1.134), $t(20)$ = 5.774, p < .001, r = .791, with similar variances $F(1, 40)$ = .078, ns.

5.3.3. *Between Groups*

After testing the results inside each group, we have computed the hypothesis tests between groups, for each measure and phase. The time spent to solve an equation in the experimental group (Mdn = 51.81) was lower than the time spent by the control group during phase 1 (Mdn = 62.13). But, a Wilcoxon rank-sum test showed that they did not differ significantly W = 286, p = .094, $r = -.256$, with similar variances $F(1, 41)$ = .01, ns. We have checked the same condition on phase 2, in which the time spent to solve an equation in the experimental group (Mdn = 82.06) was also lower than the time spent by the control group (Mdn = 97.93). And, a Wilcoxon rank-sum test showed that they differ significantly W = 326, p = .010, $r = -.391$, with similar variances $F(1, 41)$ = 3.401, ns.

For the time spent by the students to solve the steps in phase 1, on average, the control group (M = 20.18, SE = .717) spent less time than the experimental group (M = 22.11, SE = 1.134). But, based on the result of an independent t-test, this difference was not significant $t(34.01) = -1.44$, p = .0795, r = .240, with similar variances $F(1, 41)$ = 1.984, ns. During phase 2, the experimental group (M = 32.75, SE = 2.132) spent less time on steps than the control group (M = 34.71, SE = 1.910). But, based on the result of an independent t-test, this difference was not significant $t(40.28)$ = .683, p = .249, r = .107, with similar variances $F(1, 41)$ = .119, ns.

5.4. *Relationship Between Measures*

After computing the descriptive statistics and the hypothesis tests for all the measures, we have calculated the correlation between these measures[15]. We have performed this calculation for the control and experimental groups in phase 1 and phase 2. We have included 11 variables in total: (*i*) score on the pre-test (only in phase 1), (*ii*) midway-test, and (*iii*) post-test (only in phase 2), (*iv*) gain midway - pre (only in phase 1), (*v*) gain post - midway (only in phase 2), (*vi*) number of solved equations, (*vii*) number of steps, (*viii*) number of correct and (*ix*) incorrect steps, (*x*) time on the equation, and (*xi*) time on a step.

To better present the results, we have created a correlogram[16] for each group in each phase. Fig. 4 shows the correlogram of the control and experimental groups in phase 1, and Fig. 5 shows the correlogram of the control and experimental groups in phase 2. The

[15] We could not compute the significance level for Pearson's correlation because the distributions do not follow a normal distribution. Thus, we have applied the Spearman's correlation coefficient r_S, which is a nonparametric statistic to calculate the correlation between two distributions, as suggested by Field *et al.* (2012).

[16] The correlograms were computed using the *ggcorrplot* R package.

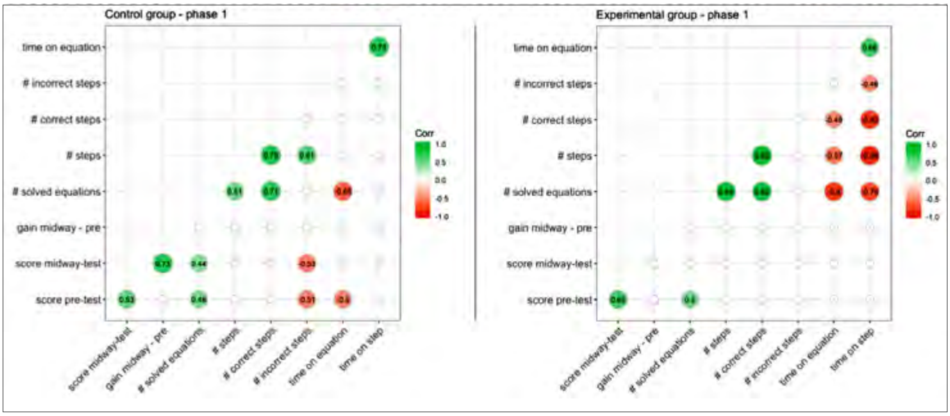


Fig. 4. Correlogram of measures from the control and experimental groups in phase 1.

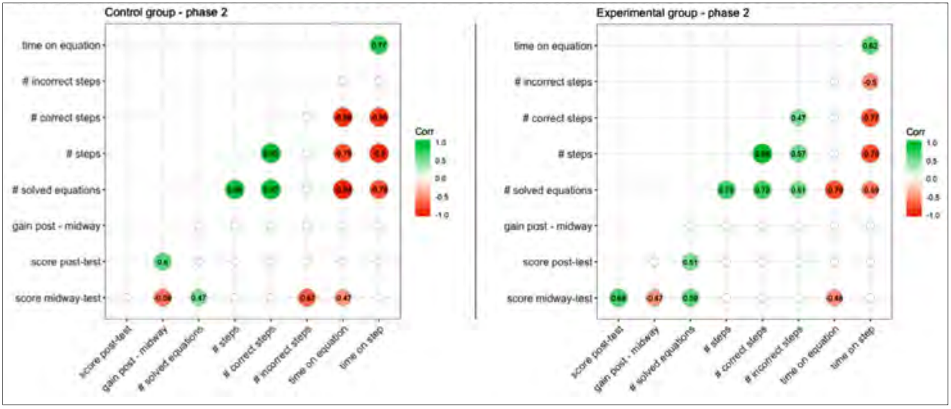


Fig. 5. Correlogram of measures from the control and experimental groups in phase 2.

calculation of the correlogram is based on a correlation matrix, calculated for each combination of measures. The blank spots presented in the correlogram represent correlations that are not statistically significant for the combination of measures. Thus, the correlograms presented in this paper show only significant correlations, i.e., *p-value* $< .05$. Also, the greener, the more positively correlated and the redder, the more negatively correlated the combinations are. Besides the color, for each significant correlation, the correlogram shows the corresponding r_s value.

6. Discussions

This section presents discussions about the results reported in Section 5, according to the collected measures, as described in Section 4.5. We have divided this section into specific subsections for each measure.

6.1. *Tests Performance*

In Section 5.1, we have presented the test scores obtained by the students in phases 1 and 2 for control and experimental groups. We have shown, with statistical significance, that students in control and experimental groups had their learning improved in phase 1 (comparing scores of the pre and midway-tests) and in the whole experiment (comparing scores of the pre and post-tests). Both results presented very large effects. This indicates that all the students learned in phase 1 and during the whole experiment using the ITS, independent of whether they used the handwriting or the typing method.

For phase 2, according to the results of the statistical tests, as we can also see in Fig. 3, the performance median of the students in the post-test was lower than in the midway-test. We have three assumptions about this result. The first one is that perhaps the students may have forgotten the content. The second one is that maybe the students may have become confused with more complex equations solved during phase 2, according to the PAT2Math ITS lessons plan structure, which gradually increases the equations' complexity according to the student's progress in the system. Moreover, the third one is that the student's performance decreased on phase 2, possibly due to the device inversion. However, we could not affirm this finding because there were no significant differences in the hypothesis tests for both groups. Although, it presented a small to medium effect on the control group and a medium effect on the experimental group.

Based on the results of the learning gain, the students in control and experimental groups achieved a higher gain on phase 1, when comparing with the gain achieved in phase 2, both groups with a very large effect. The difference between phases 1 and 2 on the experimental design was the device used to enter their equations on the ITS. It indicates that regardless of the handwriting or typing methods, students learned more in phase 1 than phase 2. During the sessions, we noticed the students' actions and the teacher's concern that the students were not satisfied using the computer's mouse for handwriting neither utilizing the tablet's keyboard to enter their solving steps in the ITS. Thus, we believe that the negative effect on learning in phase 2 is perhaps associated with device inversion.

Also, we have compared the learning gain of students between groups for different phases. In general, the students in the experimental group achieved lower gain than the students in the control group for phases 1 and 2 and the whole experiment. It means that the students who used the handwriting method had a lower performance than students who used the typing method. However, this finding was not significant for any of the tests.

We can also notice some negative mean gains during phase 2. These negative means indicate that the students achieved worse results after the ITS use sessions, perhaps because they forgot or became confused about the content. However, based on the results described for the control group (Section 5.1.1), for the experimental group (Section 5.1.2), and for between groups measurements (Section 5.1.3), it was not possible to find significantly statistical results to confirm this finding.

Based on the results of the test scores, we were able to identify that all the students, regardless of using the typing or handwriting input methods, performed equivalently. We have applied statistical tests to measure students' knowledge differences, but we could not find any remarkable difference. Although this result is not statistically significant, we believe that the equations solved by the students did not require them to insert long and complicated steps that demand multidimensional viewing.

Typically, the more complex the equations are, the more multidimensional viewing they require, for example, to solve equations with fractions. The work of Anthony *et al.* (2007b) stated that the time required for students to solve equations with fractions is almost twice when using typing *vs.* the handwriting input method. If the students had started the experiment on more complex equations, they would have entered more complicated steps containing fractions operations. It is easy to enter fractions in the student's notebooks because they can draw them as they want, using multiples lines. However, to do this on the system using the traditional keyboard, they must follow a given pattern. In the PAT2Math ITS, this pattern consists of utilizing a parenthesis structure. For example, to enter a solving step like this:

$$\frac{2+x}{3+4}-\frac{3-x}{16+8}=-10$$

the student would have to type an input like this: "$(2+x)/(3+4)-(3-x)/(16+8)=-10$" in order for the system to interpret the step. This required pattern is far away from what the students are used to write in their notebooks.

One of our hypotheses was that by allowing the student to use the handwriting with real-time recognition as an input method in step-based math ITS, receiving feedback for every solution step, the students would reason more about the content than the graphical interface, leading to reduced cognitive load and improved learning. However, based on the results we have calculated, we do not have evidence for this hypothesis. Although the students have learned during the experiment, we could not find any significant difference in learning when considering the students divided into the control and experimental groups.

Our results are in line with the findings reported by Anthony *et al.* (2007b). In their work, the authors also performed one experiment with students in the control group using the typing and the students in the experimental group using the handwriting input method. Although they have found favorable results for handwriting-based interfaces, the authors have identified no difference in student learning. We come up with two possible explanations why we could not find significant results on students learning for our work.

First, the students in this experiment did not achieve advanced content on first-degree equations, as presented in Section 4.4. Thus, we believe that on more simple equations that do not require structured input on the system, the students learn really fast to use the keyboard because they have to insert just numbers and the plus (+) or minus (−) symbols. However, this task would get more challenging for advanced equations, leading the students who would use the handwriting input method to perform better than students using the typing method. Therefore, we believe that the handwriting-based interface effectiveness may depend on the content complexity and the need for multidimensional visualization.

Second, before conducting this experiment, we have performed multiple usability evaluations on the old graphical user interface of the PAT2Math[17]. In short, we have identified that the old interface of the system had several problems, and it was inappropriate for the students to use. This old version was a stand-alone applet that followed the WIMP (windows, icons, menus, pointer) user interface style. Therefore, we have developed a new web-based interface that removes all these elements and allows the student to type in a text box freely. This developed new version was used in this experiment for the students in the control group, using the typing input method. We brought this discussion because most of the related works were also based on the WIMP style. Thus, the good results related to the handwriting-based interfaces could be related to usability problems of the other interface versions in comparison. Therefore, we hypothesized that the easier it is to use the interface, the lower the cognitive load imposed by the system, regardless of the need for multidimensional visualization. This hypothesis needs to be further investigated, and we have considered it as future work.

6.2. *Number of Solved Equations and Entered Steps*

Besides the test scores, we have also analyzed the number of equations and steps that the students have entered for different groups and phases. According to the results, for both control and experimental groups, the number of solved equations, entered steps, correct steps, and incorrect steps in phase 2 were significantly lower than in phase 1. It means that, in general, students solved more equations and entered more steps using the most suitable device for each modality, independent of the input method. The handwriting is best performed on the tablets by touchscreen, and the typing is best performed on the computer's keyboard.

When comparing the difference between groups, we were able to identify that the number of solved equations of the students in the control group was lower than the number of solved equations of the students in the experimental group. It means that students using the handwriting input could solve more equations than students using the keyboard input. Although the mean of the number of solved equations for the experimental group is greater than the control group in both phases, this finding is significant only during phase 2.

Even solving more equations, students in the experimental group entered fewer steps and fewer correct steps than students in the control group during phase 1. One possible reason for this finding is that the handwriting method may have led the students to feel more comfortable solving the equations in the way they are used to solve in their notebooks. PAT2Math does not impose any requirement about a minimum number of steps or the required mathematical operations the student must insert. However, when the student sees the input box used for typing, s/he may think that s/he has to insert every

[17] The details about these evaluations can be found here: Morais and Jaques (2013) and Morais *et al.* (2017a).

step for the system to understand her/his solution, leading to a greater number of solving steps and fewer equations solved. This assumption needs to be further investigated in future work.

Inversely, during phase 2, students in the control group entered fewer steps and fewer correct steps than students in the experimental group. Although none of these results were statistically significant, it means that even solving more equations than students using the typing input method, students that used the handwriting input method entered fewer steps and, consequently, fewer correct steps during phase 1. Nevertheless, during phase 2, the students using the typing input method entered fewer steps, fewer correct steps and solved fewer equations than students using the handwriting input method.

For the number of incorrect steps, in both phases, the students on the control group entered fewer incorrect steps than students on the experimental group. It means that students using the typing input method made fewer mistakes when entering steps than students using the handwriting input method. However, this result was not statistically significant. By analyzing the number of solved equations and entered steps in the system, we were able to identify that, in general, students using the handwriting input method solved more equations but also entered more incorrect steps than students that used the typing input method. Some of these results are not statistically significant, but it helps us understand the students' behavior. In this case, we believe that by spending more time on the typing task, the students can reflect more on the solution before entering the solving step.

6.3. *Time Spent on Equations and Steps*

This work aimed to verify the impact of the input method on a step-based ITS over the students' fluency to solve first-degree equations. Thus, we measured the time spent by the students using different input methods to solve the equations and enter the solving steps.

We have computed the time spent on the equation and steps by students in different groups and phases. According to the results, for both control and experimental groups, students on phase 2 spent significantly more time solving the equations and entering the steps than on phase 1. According to Table 1, which shows the structure of the equations solved in both phases, only a tiny quantity of the equations solved by the students in phase 2 was more difficult than phase 1, i.e., they needed more operations to be solved. Thus, although the complexity of the equations may have interfered in the time spent by the students to solve the equations, this represents a tiny number of equations solved. Therefore, we believe the main reason for this difference in the time between phases is due to the device inversion. It means that students using the handwriting input method spent less time using tablets than using the computer mouse. Also, students using the typing input method spent less time using the computer keyboard than the tablet keyboard.

When comparing the time spent between groups, we could identify that students in the experimental group spent less time to solve equations in both phases when compared to the control group. It means that the students using the handwriting input method spent less time than students using the typing input method to solve the equations, regardless of the device. This finding was statistically significant only for phase 2.

For the time spent on a step, students in the control group spent less time on a step than students in the experimental group during phase 1. Inversely, students in the experimental group spent less time on a step than students in the control group during phase 2. Nonetheless, none of these results are statistically significant. It means that students typing on the computer keyboard spent less time on a step than students handwriting on tablets. Moreover, students handwriting with the computer's mouse spent less time on a step than students typing on the tablets' keyboard. Thus, regardless of the input method, students using the computer spent less time on equations than students using the tablets.

The most remarkable result of the time measure is about the time spent solving the equations. According to our tests, we could report, with statistical significance, that students were faster to solve the equations using the handwriting than the typing input method. According to Oviatt *et al.* (2006), graphical user interfaces cause students to experience high extraneous cognitive load, directly affecting the speed at which the task is performed. Thus, by reducing the time spent to solve the equations, we can assume that we have reduced the extraneous cognitive load imposed by the graphical interface, as reported by Sweller (2010). This finding is in line with the work of Anthony *et al.* (2005, 2007b), which also reported a faster speed for handwriting-based interfaces compared to the typing method. However, Anthony *et al.* (2007b) only found this evidence for equations containing fractions. In our work, we have found this result for the whole experiment.

6.4. *Relationship between Measures*

After computing the results according to our measures, we have calculated the correlation between all the measures. The correlation helped us to find trends and understand the behavior of the students using the ITS. We have calculated the correlation for each phase.

6.4.1. *Phase 1*

During phase 1, we have observed some common correlations between the control and experimental groups. These common correlations indicate that the different input methods and devices did not influence these measures during phase 1. We can report that the higher the score on the pre-test, the higher the score on the midway-test. With R_s^2 = .28, this correlation could explain 28% of the results in the control group, and with R_s^2 = .42, this correlation could explain 42% of the results in the experimental group. Also, we have found that the greater the score on the pre-test, the more equations the students have solved.

We have also identified that the $more$ solved equations, the more steps entered by the students. This correlation is a bit obvious, but the number of steps depends on each student-solving strategy. PAT2Math ITS allows the student to solve the equations by entering the solution step-by-step or just the final answer. Thus, the student is not forced to enter more steps. Although this result is not statistically significant, when comparing different input methods, the students who used the handwriting input method entered fewer steps and solved more equations than students using the typing input method.

The more steps entered in the system or, the more solved equations, the more correct steps the student entered. When comparing the input methods, students using the handwriting input method entered, in general, fewer steps and also fewer correct steps than students using the typing input method. However, they solved more equations than students using the typing input method. None of these results were statistically significant. About time correlations, the more solved equations the student had, the less time on equation the student spent. This correlation is also a bit obvious because we have controlled the time range of the experiment. Thus, all the students had the same amount of time to use the ITS. If some of them solved a greater number of equations, they had to be faster than others. None of these results were statistically significant. Also, the more time the student spent solving the equation, the more time the student spent solving a step.

Besides the common correlations, we could find some results that were unique for each group. Thus, we are going to discuss the **striking correlations for the control group** in phase 1. The higher the score on the pre-test, the smaller the number of incorrect steps and the less time spent on equations by the students. Thus, the pre-test score can be considered an indicator of the rest of the results. Still about score tests, the higher the score on the midway-test, the greater the gain *midway-pre*, the more solved equations, and the fewer incorrect steps the students entered in the ITS. It means that solving equations on the ITS helped not just the students to achieve a higher score. Instead, it allowed the students to learn more, solve more equations, and commit fewer errors. The last correlation of the control group in phase 1 shows that the more steps entered in the system, the more incorrect steps entered by the student. This correlation was significant only for the control group. Therefore, the number of wrong steps was growing according to the number of steps for students using the typing input method.

We have also found **striking correlations for the experimental group** in phase 1. An interesting point on this is that all the correlations in the experimental group are related to time. The higher the number of correct steps or the number of steps, the less time spent by the student to solve the equation. The greater the number of incorrect steps, correct steps, number of steps, or the number of solved equations, the less time the student spent solving a step. All these correlations indicate that students who used the handwriting input method spent less time solving the steps and less time solving the equations.

6.4.2. *Phase 2*

We have also computed the correlations for phase 2, in which the students switched devices to enter equations in the system. Again, we have identified some **common correlations between the control and experimental groups** for phase 2. These correlations that are common between groups indicate that the input method did not affect the results.

The higher the score on the midway-test, the smallest the *post-midway* gain, the more solved equations, and the less time spent on the equation during phase 2. The more solved equations, the higher the number of steps and correct steps, and the less time spent on the equation and step. The more time spent on the equation, the more time spent on a step. Also, the higher the number of correct steps, or the number of steps, the less time spent on a step. And, the higher the number of steps, the greater the number of correct steps.

We have found **striking correlations for the control group** in phase 2. The higher the score on the post-test, the higher the *post-midway* gain. The higher the score on the midway-test, the smaller the number of incorrect steps. The higher the number of correct steps or the number of steps, the less time spent on the equation by the student. We have also found **striking correlations for the experimental group** in phase 2. The higher the score on the midway-test, the greater the score on the post-test. The more solved equations, the higher the score on the post-test. The higher the number of correct steps, the number of steps, or the number of solved equations, the higher the number of incorrect steps. The higher the number of wrong steps, the less time the student spends to solve a step.

6.4.3. *Between Phases*

We have compared the correlation matrices between phases, considering only significant correlations, aiming to identify the correlations regardless of input methods or devices. We could observe that, in general, the number of steps, correct steps, and solved equations grows linearly. The number of the solved equations depends on the score on the first test of each phase, i.e., the pre-test for phase 1 and midway-test for phase 2. Also, the time spent solving an equation is smaller for the students that solved more equations. Furthermore, the time to solve an equation is dependent on the time to solve the steps.

7. Conclusions

This paper describes an experiment with a math ITS, PAT2Math, which was integrated with a plugin that allows students to insert equations into the system through their handwriting. In this version, the student can see the recognized handwriting in real-time and receives specialized step-based feedback from the ITS to solve first-degree equations. The research hypothesis of this work is that this combination of real-time handwriting recognition with step-by-step guidance from an ITS would provide a more natural input data approach compared to the typing method. Thus, the students would have more fluency to solve the equations, reducing the extraneous cognitive load imposed by the typing method, leading to improved learning.

In the typing method, the students have to memorize the text-like patterns to insert the equations in the system. We have replaced the traditional typing input box with a handwriting area that recognizes the student's handwriting in real-time. We have conducted an experiment in which 55 students were randomly distributed to control and experimental groups. Whereas the students in the control group used the typing method, the students in the experimental group used the handwriting approach.

The experiment comprised two phases. The students in the control group used the computer's keyboard for phase 1 and the tablet's keyboard for phase 2 to type the equations on the ITS. The students in the experimental group used the tablet's touchscreen for phase 1 and the computer's mouse for phase 2 to handwriting the equations. By collecting the students' data about their interaction while using the ITS and the performance scores from knowledge tests for each phase, we were able to compute the learning gain, the number of solving equations, and the time spent by the students to solve the equations.

We have found that students in both groups significantly learned by using the ITS. However, there was no difference in the performance test scores between groups for both phases. It indicates that the different input methods did not impact students learning. Students in both groups achieved significantly lower gains in phase 2 than in phase 1. Although this result was not statistically significant, it indicates that the handwriting input method with the mouse on the computer and the typing input method in the tablet keyboard may negatively affect the students' learning.

We also have compared the number of solved equations between phases and groups. In phase 1, the students solved more equations and entered more steps, correct steps, and incorrect steps than phase 2. It indicates that the handwriting input method was a better strategy when students used tablets or touch screen devices instead of the computer mouse. The same is true for the typing input method, which is better on a computer keyboard than a tablet keyboard. In general, we have noticed that students using the handwriting input method solved more equations than the students using the typing input method. However, conversely, they also entered more incorrect steps than students who typed.

Finally, about the time measure, in general, the students spent more time in phase 2 than in phase 1 to solve the equations and to enter the steps in the system. Although a small set of the equations in phase 2 was more difficult them the equations solved in phase 1, this finding may indicate that the handwriting input method is more efficient on tablet devices, and the typing input method is more productive on the computer keyboard. Comparing input methods, the students using the handwriting method spent less time solving the equations than students using the typing method.

Thus, we can conclude that the handwriting input method allows the student to be faster in solving equations. We can attribute this speed difference to a reduced cognitive load interface, which helps students insert the equations more fluently using their handwriting. Although this finding is related to our hypothesis, we cannot assume this measure is enough to define if the handwriting input method can impact the students learning.

In general, we have found that the input method did not impact the students learning. This finding is also in line with related studies in this area. On the other side, the device used to insert the equations, when used with not ideal input methods, had a more significant effect, negatively affecting the students' performance. Thus, we can conclude that the handwriting input method is not able to impact the learning of the students using math step-based ITS to solve not too complex first-degree equations.

7.1. *Limitations*

This section highlights some limitations of this work. We have conducted the experiment with a small number of students from the same private school from southern Brazil, which does not allow us to generalize our findings to the global population.

The midway-test was conducted differently from the pre and post-tests. Whereas the pre and post-tests were paper-based, the midway-test was a computer-based test. We have decided not to expose the students to three sequential performance tests, in the same manner, in a short period. We have controlled all the variables not to cause any result difference. However, this computer-based test could have motivated the students once it was their first time solving a performance test on the computer.

Finally, the experiment was conducted during six ITS usage sessions, which did not allow the students to go further on the content. In the PAT2Math ITS, the more equations the student solves, the more complex the equations become. As the experiment was not long enough to achieve the hardest content, students did not solve complex equations that required more text-like patterns to be typed, for typing input method only. We believe that to solve equations with more operations, the students using the typing input method would spend much more time than students using the handwriting input method to solve the same equations. This difference could have also impacted the test scores.

7.2. *Future Work*

The students using the typing input spent more time solving the equations and entered fewer wrong steps than the students using the handwriting method. Thus, as future work, we plan to investigate whether the extra time required from students during typing leads them to more reflection before entering the solving step. Also, the students using the handwriting method solved more equations and entered fewer steps than students who used typing. Thus, we plan to research whether the students using the typing have entered more steps due to the belief that the ITS only recognizes the solving when entered step-by-step.

We also plan to verify whether solving more complex equations, i.e., first-degree equations containing more fractions during solving, could affect students' fluency using the typing input method due to the multidimensional viewing requirement. When using a typing-based interface, the students must insert the solving steps following a set of patterns. In the case of fractions, this text-like equation pattern would be challenging for the students to remember. Finally, we have seen many works reporting favorable results to the handwriting input method had the previous graphical interface version based on the WIMP pattern. Thus, we plan to investigate whether the usability of the interfaces could impact when testing different input methods.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, FAPERGS (granting 17/25510001203-8) and CNPq (granting 306005/2020-4) from Brazil.

References

Alvarado, C., Kearney, A., Keizur, A., Loncaric, C., Parker, M., Peck, J., Sobel, K., Tay, F. (2015). LogiSketch: A Free-Sketch Digital Circuit Design and Simulation SystemLogiSketch. In: *Hammond, Tracy and Valentine, Stephanie and Adler, Aaron and Payton, Mark (Eds.) The Impact of Pen and Touch Technology on Education. Human-Computer Interaction Series*. Springer, Cham, pp. 83–90. https://doi.org/10.1007/978-3-319-15594-4_8

Anthony, L., Yang, J., Koedinger, K.R. (2005). Evaluation of Multimodal Input for Entering Mathematical Equations on the Computer. In: *van der Veer, Gerrit C. and Gale, Carolyn (Eds.) Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’05. Association for Computing Machinery, New York, NY, USA, pp. 1184–1187. https://doi.org/10.1145/1056808.1056872

Anthony, L., Yang, J., Koedinger, K.R. (2007a). Adapting Handwriting Recognition for Applications in Algebra Learning. In: *Friedland, Gerald and Hürst, Wolfgang and Knipping, Lars and Mühlhäuser, Max (Eds.) Proceedings of the International Workshop on Educational Multimedia and Multimedia Education*. Emme ’07. Association for Computing Machinery, New York, NY, USA, pp. 47–56. https://doi.org/10.1145/1290144.1290153

Anthony, L., Yang, J., Koedinger, K.R. (2007b). Benefits of handwritten input for students learning algebra equation solving. *Frontiers in Artificial Intelligence and Applications*, 158, 521–523. https://doi.org/10.5555/1563601.1563683

Anthony, L., Yang, J., Koedinger, K.R. (2012). A paradigm for handwriting-based intelligent tutors. *International Journal of Human-Computer Studies*, 70(11), 866–887. https://doi.org/10.1016/j.ijhcs.2012.04.003

Barreto, L., Taele, P., Hammond, T. (2016). A Stylus-Driven Intelligent Tutoring System for Music Education Instruction. In: *Hammond, TracyandValentine, StephanieandAdler, Aaron(Eds.) Revolutionizing Education with Digital Ink: The Impact of Pen and Touch Technology on Education*. Springer, Cham, pp. 141–161. https://doi.org/10.1007/978-3-319-31193-7_10

Cheema, S., LaViola, J. (2012). PhysicsBook: A Sketch-Based Interface for Animating Physics Diagrams. In: *Duarte, Carlos and Carriço, Luís and Jorge, Joaquim and Oviatt, Sharon and Gonçalves, Daniel (Eds.) Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*. IUI ’12. Association for Computing Machinery, New York, NY, USA, pp. 51–60. https://doi.org/10.1145/2166966.2166977

Cheema, S., LaViola Jr., J.J. (2018). Using Animation to Enrich Learning Experience in Sketch-Based Physics Tutoring Systems. In: *Kapros, Evangelos and Koutsombogera, Maria (Eds.) Designing for the User Experience in Learning Systems*. Springer, Cham, pp. 201–227. https://doi.org/10.1007/978-3-319-94794-5_10

Cummmings, D., Vides, F., Hammond, T. (2012). I Don’t Believe My Eyes! Geometric Sketch Recognition for a Computer Art Tutorial. In: *Singh, Karan and Kara, Levent B. (Eds.) EUROGRAPHICS Symposium on Sketch-Based Interfaces and Modeling*. SBIM ’12*: Vol.* 12. The Eurographics Association, Goslar, DEU, pp. 97–106. https://doi.org/10.2312/SBM/SBM12/097-106

Dixon, D., Prasad, M., Hammond, T. (2010). ICanDraw: Using Sketch Recognition and Corrective Feedback to Assist a User in Drawing Human Faces. In: *Mynatt, Elizabeth and Fitzpatrick, Geraldine and Hudson, Scott and Edwards, Keith and Rodden, Tom (Eds.) Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’10. Association for Computing Machinery, New York, NY, USA, pp. 897–906. https://doi.org/10.1145/1753326.1753459

Elliott, E.S., Dweck, C.S. (1988). Goals: An approach to motivation and achievement. *Journal of personality and social psychology*, 54(1), 5–12. https://doi.org/10.1037/0022-3514.54.1.5

Field, A., Miles, J., Field, Z. (2012). *Discovering statistics using R* (1st ed.). SAGE Publications Ltd, London.

Glaser, R. (1976). Components of a psychology of instruction: Toward a science of design. *ReviewofEducational Research*, 46(1), 1–24. https://doi.org/10.3102/00346543046001001

Graesser, A.C., Hu, X., Nye, B., Sottilare, R. (2016). Intelligent tutoring systems, serious games, and the Generalized Intelligent Framework for Tutoring (GIFT). *Harold F. O'Neil, Eva L. Baker, Ray S. Perez (Eds.) Using games and simulation for teaching and assessment*, 58–79.

Hilton, E., Williford, B., Li, W., Hammond, T., Linsey, J. (2019). Teaching Engineering Students Freehand Sketching with an Intelligent Tutoring System. In: *Hammond, Tracy and Prasad, Manoj and Stepanova, Anna (Eds.) Inspiring Students with Digital Ink: Impact of Pen and Touch on Education*. Springer, Cham, pp. 135–148. https://doi.org/10.1007/978-3-030-17398-2_9

Jaques, P.A., Seffrin, H., Rubi, G., de Morais, F., Ghilardi, C., Bittencourt, I.I., Isotani, S. (2013). Rule-based expert systems to support step-by-step guidance in algebraic problem solving: The case of the tutor PAT2-Math. *Expert Systems with Applications*, 40(14), 5456–5465. https://doi.org/10.1016/j.eswa.2013.04.004

Kang, B., Kulshreshth, A., LaViola, J.J. (2016). AnalyticalInk: An Interactive Learning Environment for Math Word Problem Solving. In: *Nichols, Jeffrey and Mahmud, Jalal and O'Donovan, John and Conati, Cristina and Zancanaro, Massimo (Eds.) Proceedings of the 21st International Conference on Intelligent User Interfaces*. IUI '16. Association for Computing Machinery, New York, NY, USA, pp. 419–430. https://doi.org/10.1145/2856767.2856789

Kang, B., LaViola Jr., J.J., Wisniewski, P. (2017). Structured Input Improves Usability and Precision for Solving Geometry-Based Algebraic Problems. In: *Mark, Gloria and Fussell, Susan and Lampe, Cliff and m.c. schraefel, and Pablo Hourcade, Juan and Appert, Caroline and Wigdor, Daniel (Eds.) Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Association for Computing Machinery, New York, NY, USA, pp. 4692–4702. https://doi.org/10.1145/3025453.3025468

Koedinger, K.R., Kim, J., Jia, J.Z., McLaughlin, E.A., Bier, N.L. (2015). Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC. In: *Kiczales, Gregor and Russell, Daniel M. and Woolf, Beverly (Eds.) Proceedings of the Second (2015) ACM Conference on Learning @ Scale*. L@S '15. Association for Computing Machinery, New York, NY, USA, pp. 111–120. https://doi.org/10.1145/2724660.2724681

Koedinger, K., Anderson, J., Hadley, W., Mark, M. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education (IJAIED)*, 8(1), 30–43.

Laviola, J.J. (2007). Advances in mathematical sketching: Moving toward the paradigm's full potential. *IEEE Computer Graphics and Applications*, 27(1), 38–48. https://doi.org/10.1109/MCG.2007.2

Lee, C., Jordan, J., Stahovich, T.F., Herold, J. (2012). Newtons Pen II: An intelligent, sketch-based tutoring system and its sketch processing techniques. In: *Singh, Karan and Kara, Levent B. (Eds.) EUROGRAPHICS Symposium on Sketch-Based Interfaces and Modeling*. SBIM '12. The Eurographics Association, Goslar, DEU, pp. 57–65. https://doi.org/10.2312/SBM/SBM12/057-065

Morais, F., Jaques, P. (2013). Avaliação de usabilidade do Sistema Tutor Inteligente PAT2Math. *Revista de Novas Tecnologias na Educação -RENOTE*, 11(3). https://doi.org/10.22456/1679-1916.44929

Morais, F., Jaques, P. (2017). PAT2Math + Handwriting: Evoluindo Sistemas Tutores de Matemática com reconhecimento da escrita à mão. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação -SBIE)*, 28(1), 1237. https://doi.org/10.5753/cbie.sbie.2017.1237.

Morais, F., Schaab, B., Jaques, P. (2017a). The think aloud method for qualitative evaluation of an intelligent tutoring system interface. In: *2017 Twelfth Latin American Conference on Learning Technologies (LACLO)*, pp. 1–8. https://doi.org/10.1109/LACLO.2017.8120904

Morais, F., Bittencourt, I.I., Isotani, S., Jaques, P.A. (2017b). The Use of Handwriting Input in Math Tutoring Systems: An Use Case with PAT2Math. In: *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*, pp. 44–46. https://doi.org/10.1109/ICALT.2017.142

Mouchère, H., Viard-Gaudin, C., Zanibbi, R., Garain, U. (2014). ICFHR 2014 Competition on Recognition of On-Line Handwritten Mathematical Expressions (CROHME 2014). In: *Gatos, Basilis and Katsouros, Vassilis and Pratikakis, Ioannis (Eds.) 2014 14th International Conference on Frontiers in Handwriting Recognition*, pp. 791–796. https://doi.org/10.1109/ICFHR.2014.138

Mouchère, H., Viard-Gaudin, C., Zanibbi, R., Garain, U. (2016). ICFHR2016 CROHME: Competition on Recognition of Online Handwritten Mathematical Expressions. In: *Suen, Ching Y. and Liu, Cheng-Lin and Chen, Youbin and Pal, Umapada and Cheriet, Mohamed and Marinai, Simone (Eds.) 2016 15th In-*

ternational Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 607–612. `https://doi.org/10.1109/ICFHR.2016.0116`

Oviatt, S., Arthur, A., Cohen, J. (2006). Quiet Interfaces That Help Students Think. In: *Wellner, Pierre and Hinckley, Ken (Eds.) Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology*. UIST '06. Association for Computing Machinery, New York, NY, USA, pp. 191–200. `https://doi.org/10.1145/1166253.1166284`

Pacheco-Venegas, N.D., López, G., Andrade-Aréchiga, M. (2015). Conceptualization, development and implementation of a web-based system for automatic evaluation of mathematical expressions. *Computers & Education*, 88, 15–28. `https://doi.org/10.1016/j.compedu.2015.03.021`

Phon-Amnuaisuk, S., Omar, S., Au, T.-W., Ramlie, R. (2015). Mathematics Wall: Enriching Mathematics Education Through AI. In: *Tan, Ying and Shi, Yuhui and Buarque, Fernando and Gelbukh, Alexander and Das, Swagatam and Engelbrecht, Andries (Eds.) Advances in Swarm and Computational Intelligence*. Springer, Cham, pp. 309–317. `https://doi.org/10.1007/978-3-319-20469-7_33`

Read, J.C., MacFarlane, S.J., Casey, C. (2000). Where's the 'm' on the keyboard, mummy. In: *Paper presented at Womens' Engineering Society*. Preston, Lancs.

Read, J.C., MacFarlane, S.J., Casey, C. (2001). Can natural language recognition technologies be used to enhance the learning experience of young children. In: *Paper presented at Computers and Learning*. Warwick, UK.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review*, 22(2), 123–138.

Sweller, J., van Merriënboer, J.J., Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review (2019)*, 31(2), 261–292. `https://doi.org/10.1007/s10648-019-09465-5`

Taele, P., Hammond, T. (2009). Hashigo: A next-generation sketch interactive system for japanese kanji. In: *Haigh, Karen and Rychtyckyj, Nestor (Eds.) Proceedings of the Twenty-First Innovative Applications of Artificial Intelligence Conference (2009)*, pp. 153–158.

Taele, P., Peschel, J., Hammond, T. (2009). A sketch interactive approach to computer-assisted biology instruction. In: *Hammond, Tracy Anne (Eds.) 2009 Intelligent User Interfaces Workshop on Sketch Recognition IUISR* (Vol. 9), pp. 3999–4005.

Thompson, R., Tanimoto, S., Berninger, V., Nagy, W. (2016). Design Studies for Stylus and Finger-Based Interaction in Writing Instruction on Tablets. In: *Hammond, Tracy and Valentine, Stephanie and Adler, Aaron (Eds.) Revolutionizing Education with Digital Ink: The Impact of Pen and Touch Technology on Education*. Springer, Cham, pp. 51–69. `https://doi.org/10.1007/978-3-319-31193-7_4`

Tran Minh Khuong, V., Phan, M.K., Nakagawa, M. (2019). Interactive User Interface for Recognizing Online Handwritten Mathematical Expressions and Correcting Misrecognition. In: *Sidere, Nicolas and Siddiqi, Imran Ahmed and Ogier, Jean-Marc and Djeddi, Chawki (Eds.) 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)* (Vol. 2), pp. 26–30. IEEE. `https://doi.org/10.1109/ICDARW.2019.10034`

Valentine, S., Lara-Garduno, R., Linsey, J., Hammond, T. (2015). Mechanix: A Sketch-Based Tutoring System that Automatically Corrects Hand-Sketched Statics Homework. In: *Hammond, Tracy and Valentine, Stephanie and Adler, Aaron and Payton, Mark (Eds.) The Impact of Pen and Touch Technology on Education*. Springer, Cham, pp. 91–103. `https://doi.org/10.1007/978-3-319-15594-4_9`

Vanlehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education (IJAIED)*, 16(3), 227–265.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221. `https://doi.org/10.1080/00461520.2011.611369`

Vuong, B.-Q., He, Y., Hui, S.C. (2010). Towards a web-based progressive handwriting recognition environment for mathematical problem solving. *Expert Systems with Applications*, 37(1), 886–893. `https://doi.org/10.1016/j.eswa.2009.05.091`

Wang, G., Bowditch, N., Zeleznik, R., Kwon, M., LaViola, J.J. (2016). A Tablet-Based Math Tutor for Beginning Algebra. In: *Hammond, Tracy and Valentine, Stephanie and Adler, Aaron (Eds.) Revolutionizing Education with Digital Ink: The Impact of Pen and Touch Technology on Education*. Springer, Cham, pp. 91–102. `https://doi.org/10.1007/978-3-319-31193-7_6`

Wicki, W., Lichtsteiner, S.H., Geiger, A.S., Müller, M. (2014). Handwriting fluency in children. (Vol. 73(2)). Verlag Hans Huber, pp. 87–96. `https://doi.org/10.1024/1421-0185/a000127`

F. Morais graduated in Computer Science (2017), with an exchange period at the University of Missouri (EUA), has a Masters in Applied Computing (2019) from Universidade do Vale do Rio dos Sinos (UNISINOS), and is currently a Ph.D. student with the Graduate Program in Applied Computing (PPGCA) at UNISINOS, holding a full scholarship financed by the CAPES agency from Brazil. Felipe has several national and international publications in renowned conferences and journals in Computers in Education. He is also part of the scientific community, being on the editorial board of the IEEE Technical Committee on Learning Technology and participating on the program committee of renowned international conferences such as EC-TEL and LACLO. Felipe's research interests are associated with Educational Data Mining, Learning Analytics, Affective Computing, and Intelligent Tutoring Systems.

P.A. Jaques is a fellow of a CNPq research productivity scholarship. She is currently a professor and researcher at the Program of Graduate Studies in Applied Computing (PPGCA) at Universidade do Vale do Rio dos Sinos (Unisinos). She published several articles in national and international journals and papers in conference proceedings, besides editing books and writing some chapter books. Patricia is also on the Editorial Board of the Springer Journal on Multimodal User and Associate Editor of IEEE Transactions on Learning Technologies (TLT) and Frontiers in Artificial Intelligence (Section AI for Human Learning and Behavior Change), besides participating of the program committee of several renowned national and international conferences on Artificial Intelligence and Computers in Education, such as ITS, ACII and others. Patricia was also the coordinator of international research cooperation Projects with France CAPES / COFECUB PRAIA (2006–2010), coordinator of the project STIC-AMSUD ACAI (2011–2012), and STIC-AMSUD EMPATIA project (2018–2019). Patricia Jaques works in the research areas Artificial Intelligence and Affective Computing, with application in Education. Her research involves detecting students' emotions and other affective states and developing systems that respond to those emotions to promote more learning and well-being.