Applying Item Response Theory to Evaluate Instruments of Introductory Programming Skills Measurement

Jucelio S. Santos*, Wilkerson L. Andrade*, João Brunet*, Monilly Ramos Araujo Melo*
*Federal University of Campina Grande (UFCG), Campina Grande, Brazil
jucelio@copin.ufcg.edu.br, {wilkerson, joao.arthur}@computacao.ufcg.edu.br, monillyramos@gmail.com

***Abstract*—This Research-to-practice Full Paper presents an exploratory and preliminary investigation on the reliability and validity of instruments for measuring introductory programming skills. Our data set consists of the performance of 30 students who participated in the experiment of a Brazilian university. We provide participants with instructional material, practical problems and solutions based on different experimental conditions. The results suggest that the instruments have a good internal consistency index and items with excellent psychometric properties. In addition, some initial evidence to suggest practicing each of the four programming skills. We found that participants who received practice in all four skills obtained a better estimate of study material and Assessment, especially for the most advanced knowledge of programming introduction.**

I. Introduction

Programming is a complicated discipline to learn and apply, as it has several concepts and structural regulations that need to be understood before achieving anything relevant. Current introductory instructions fail to identify, structure, and sequence the many skills involved in [1] programming.

With that in mind, Xie et al. [2] created an instructional theory for introductory programming skills. This theory proposes the development of specific and incremental skills to avoid overloading students. These skills include: reading semantics, writing semantics, reading models, and writing models. First, students develop language-related skills by teaching tracking and then writing the correct syntax. Then, students start using the models, understanding common patterns in the code, and writing programs that use those models. The most salient part of this theory is not necessarily the four specific skills. Moreover, yes, the emphasis on being rigorous and explicit in the way skills are defined and ordered in instructions.

Based on this theory, the authors [2] have created learning materials for a subset of programming concepts, which include instructional content, hands-on exercises with feedback, and a post-test covering the four skills. They then conducted an exploratory mixed-method assessment of this curriculum with two groups of beginning programmers to explore the theory's validity more broadly. They compared exercise completion rates, error rates, ability to explain code, and involvement in learning the program. The study indicated that the teaching of skills resulted in an improvement in the rate of completion of practical exercises, a decrease in the rate of errors, and post-test comprehension.

However, Xie et al. [2] did not analyze the psychometric properties of the items of the Practical Instrument and Assessment Instrument to determine the construct's validity. Moreover, they did not measure the results of students with different prior knowledge, learning contexts, and motivations. In other words, they did not measure the validity of the instrument. In this article, we aim to answer the following research questions: Are the instruments reliable? Do the instruments have evidence of validity?

To answer these questions, we use the Classical Test Theory (CTT), and the Item Response Theory (IRT) [3] [4] to analyze [1]: I) reliability of the instruments through the internal consistency and the psychometric properties of the items. Also, ii) estimate the skills of individuals as a way of verifying whether the instruments can distinguish groups. We investigated whether the explicit statement and the practical provision of the ability to read semantics and reading models improve student performance in semantics and writing models.

Our data set consists of the performance of 30 students who participated in an introductory programming course at a Brazilian university. we developed/conducted our study in the context of an introductory programming course and using the Python language. We provide participants with study material with instructions, practical problems, and solutions based on different experimental conditions. Then, we evaluated the participants to investigate how much they learned. We analyzed the data by the IRT with the aid of the R-Studio tool through R packages for analysis and adjustment of the two-parameter logistic model (2PL).

We have organized this article as follows. In Section II, we report related work. In Section III, we describe skill theory and measurement models. In Section IV, we discuss the method used in this study. In section V, we offer the results and their discussions in section VI. Finally, in Section VII, we present the conclusions and suggestions for future work.

II. Related Work

Various methods, processes, approaches, and instruments support professors and students to improve the evaluation

[1]CCT and IRT are two references used to construction, validation, and evaluation of instruments [5]. They are theories that evaluate cognitive constructions. They are not contrary theories, but the IRT is a complement to the limitations of the TCR.

process in teaching programming. As the paper of Lister et al. [6] reports a project that explored students' skills in basic syntax, tracking code, comprehension code, and writing code, seeking to establish the relationships between these skills, following the same principle as Xie et al. [2].

However, these works have no construct validity about their instruments. For CCT and IRT, the trust and validation of an instrument are critical/crucial steps as they will be the ones that will determine the level of scientific evidence that that Instrument has on a given construct [3] [4].

In Computing [7] [8] [1], the practice of these theories cited has become usual in the scientific community. Studies have proven the validity of instruments that measure individual skills.

Bergersen et al. [7] have developed/validated an instrument that measures programming skills. The Instrument had satisfactory (internal) psychometric properties. They correlated with external variables, according to theoretical expectations. Araújo et al. [8] explored the evaluation of computational thinking in introductory programming courses using the Bebras test. They conducted a preliminary study on the use of TRI to improve the selection of questions and the instrument design. In Wang et al. [1], they used IRT in the process of modeling students' learning results in a programming language course.

In this study, we will explore CTT and IRT on instruments to measure introductory programming skills created to test the Theory of Instruction for Introductory Programming Skills [2]. We will consider the same experimental conditions to assess the confidence of the instruments, as well as their level of validity. In addition to the results indicating the right internal consistency of the instruments and items with excellent psychometric properties, the study also pointed out that the performance of educated participants who practiced the ability to read semantics and reading models surpassed the writing models and semantic code when compared to participants in the study. Thus, this study contributes to the scientific community by indicating strong evidence between the skills of tracking and writing code, as explored in other studies [9] [10].

III. Background

This section describes Theory of Instruction for Introductory Programming Skills [2], CTT and IRT [3] [4].

A. Theory of Instruction for Introductory Programming Skills

The theory identifies four distinct and incremental skills involved in introductory programming teaching. These skills are:

1) Reading Semantics: refers to the skill to accurately track code and predict the effect of syntax on program behavior. Reading semantics requires the student to track the code and does not require knowledge of other skills. This practice consists of examining fixed code questions in which the student determines the intermediate and final states of the program for a predefined piece of code.

2) Writing Semantics: after practicing reading semantics and receiving feedback in the form of the correct solution and an explanation, students begin to learn to write the correct syntax. Writing semantics refers to the translating of unambiguous natural language descriptions of language constructs into syntax that will compile and execute as expected.

3) Reading Templates: after learning to read and write semantics for a new construct, a learner then transitions to how to use templates of common code use patterns to apply knowledge of this construct. The reading template refers to the skill of identifying reusable abstractions of programming knowledge (which we will refer to as templates) and mapping them to an objective.

In total, the theory taught four templates: Variable swap, Digit processing, Float equality, and Max/min. The instructions for the reading template usually started with an example or visualization to make the model objects and the steps more concrete.

4) Writing Templates: requires a learner to start with a problem description that contains ambiguity, identify a template that they could use to solve the problem, and implement each component of the template in code.

While the writing semantics instruction specifies rules to avoid syntactic errors, the writing template instruction specifies rules to avoid logic errors that would be syntactically correct but would result in code that did not work with the template specification.

B. Classical Test Theory

The CTT considers the total score of an instrument as the primary measure of the performance evaluation of an individual. Also, CTT uses norms to interpret the scores of an instrument, and these norms reference the interpreters and classify the scores, for example, to position a position of an individual that is not measured by instruments or two subject score instruments [4].

The CCT is associated with the difference between the untested individual's score and the actual value of this score [11]. Already valid, the CCT proposes to verify whether the construct measures or should measure [5].

Through the score, it is possible to use some measures to evaluate the quality of the items and the Instrument, such as the biserial point correlation coefficient and Cronbach's alpha coefficient.

1) Biserial Point Coefficient: In educational tests, it is possible to calculate the correlation coefficient between two variables, one numerical and one nominal categorical. In this case, the categorical variable has only two possible values (right/wrong); one such variable is called a dichotomous. Then, to calculate the correlation between this variable and another variable (numeric), proceed with the calculation of Pearson's coefficient in the usual way, given the assumptions of normality of the sample, which is called the biserial point correlation coefficient [12].

The biserial point coefficient estimates, which are the items in which, if the evaluated subject hits this item, is more likely

to achieve better results in an instrument. This measure works by the following characteristic, the higher the coefficient, the stronger the correlation of that item with the score and indicates that item is essential for the total result of the Instrument. This measure ranges from -1 to 1, and the closer to 1, the more discriminative the item [4].

2) Cronbach's Alpha Coefficient: After building the Instrument, one of the most efficient ways, in terms of time and cost, is to check its reliability. The internal consistency measures the reliability of the Instrument. The internal consistency consists of examining the homogeneity of the items that make up the Instrument, that is, verifying the magnitude of the relationships between the items and the total score. We can calculate internal consistency from the overall score of the Instrument and the score for each item.

We calculated this consistency using Cronbach's alpha coefficient ranging from 0 to 1; the closer to 1 indicates that the Instrument has adequate internal consistency [13]. Values closer to 1 indicate that the Instrument has adequate internal consistency. Values between 0.70 and 0.80 are considered acceptable, but with caveats. When the values are below 0.70, it means that the items that make up the Instrument need to be reassessed by the researcher [4].

C. Item Response Theory

We can use the IRT in the elaboration of educational assessment tests, item calibration (characterization of items by numerical parameter values), and other processes related to testing development. IRT, it is possible to adjust the data for the model. Thus, different people or the same person at different times may have their skills compared to everyday test items because they use parameters that are statistically measured regardless of the sample used [3].

In IRT, a set of hypothetical factors or variables can predict the individual's behavior in an item. Also, the dependency between behavior and skill may be related to a growing monotonous mathematical function, whose graph is called the Item Characteristic Curve (ICC) [5] [14].

1) Item Characteristic Curve: The ICC provides information on the probability of each getting the item right [12]. Different mathematical models may be used depending on the number of parameters involved, dimensionality, or type of items present in the Instrument. In this paper, we considered the two-parameter one-dimensional logistic model: threshold and slope.

The threshold of an item refers to the skill required for an individual with a given probability of hitting the item, calculated from the probability of hitting the item by chance, ranging from -4 (easy items) to +4 (hard items), passing through the value 0 (median items). In turn, the slope of the item refers to the inclination of the ICC. It describes how individuals of different abilities differ in the probability of hitting the item, the power to specify subjects with proximate magnitudes in the latent trait. to which it refers, ranging from 0 (non-discriminative) to 4 (extremely discriminative) [15].

Through the multiplicity of procedures of specialized computer programs can obtain the parameters of the item. These programs use nonlinear mathematical functions, such as logarithmic functions, which produce ICC, and graphical representations of mathematical functions that relate the probability of item response to latent trait level or skill [4].

2) Latent Variable Estimate: Previously introduced, 2PL reproduces a scale called latent trait or skill. The generated level is standardized (mean = 0 and SD = 1), and, as observed to the parameter b metric, in theory, this scale may range from -4 to +4. Thus, the scores estimate by the IRT using an estimation method [16].

In this study, we used the Expected A Posteriori (EAP) [17] [18]; each person gets given the score that best identifies their skill on the scale. The EAP procedure estimates the capacity of an examiner. This is a mean of a posteriori distribution and the standard error after application of the Instrument and depends on Item Information Function (IIF) and its parameters.

3) Item Information Function: The IIF analyzes how much an item contains psychometric information for skill measurement. Being statistically defined as the amount of psychometric information an item contains at all points along the continuum of the latent trait, it represents [5].

The IIF is a powerful tool for item analysis, allowing us to know not only how much information an item accumulates at a given value of θ, but also at what value of θ the item has the most amount of information. IIF has been the most commonly used method of item analysis by test builders today [12].

IV. Method

This section presents the research planning we conducted in the 2019 school year. This study sought to understand how instruments reflect the theory cited by Xie et al. [2] in reliability and evidence of validity.

A. Participants

We selected 30 beginning students of the Computer Science course at a Brazilian university to participate in this experiment. We also allocated participants to two study groups called the experimental group and the control group. It is worth mentioning that the sample selection procedure was characterized as available, estimating the ability of the participants present on the day.

We divided the participants of the experiment according to a subjective analysis of confidence in Python (Survey available in https://github.com/codeandcognition/archive-2018csexie). To avoid mistakes in this division - a group receiving participants with excellent performance, and others don't - We applied a survey before the study. We adopted a seven-point Likert scale assessment model in the research, and we represented the answer to each question is as follows: 1 (by no means confident) to 7 (absolutely confident). Based on the marks awarded, we averaged the answers to each question.

Then we test the normality of the data and get the value p = 0.017. Thus, we can conclude that the sample does not follow a normal distribution with a confidence level of 95%. Since the

data do not follow a normal distribution, we apply the Mann-Whitney test to check for statistically significant differences between the performance of the groups, as shown in Table I.

TABLE I
TRUST IN PYTHON ANALYSIS, BETWEEN GROUPS

Null hypothesis	Control Group		Experimental Group		P-value
	Median	Standard Deviation	Median	Standard Deviation	
There is significance in the difference between groups	3.45	1.73	3.70	1.44	0.047

The test shows that this difference between groups is not significant. We consider as a null hypothesis, the level of confidence in Python in the experimental group is different when compared to the control group. In the pre-survey, according to the p-value of 0.047, it shows that we distributed individuals evenly among groups, that is, without having more qualified individuals in one group when compared to the other.

B. Instruments

The instruments are: i) Study Material; And, ii) Assessment. These instruments work on a subset of programming concepts, including data types, variables, arithmetic operators, printing statements, relational operators, and conditional statements [2].

The Study Material includes instructional content and practical exercises with feedback. The Assessment content evaluative practices.

1) Study Material: Due to the experimental conditions of this study, the Study Material has two versions. The versions of the instruments, used in the control group and the experimental group, are available in https://github.com/codeandcognition/archive-2018csexie.

The material used in the experimental group labels and provides practice for each skill. While the control condition acquired practice only in Writing Semantics and Writing Templates, to balance the amount of learning received, we provided the group with additional practical writing control when compared to the experimental group, which had reading and (less) writing practice. Therefore, although the type of training varies between groups, we try to balance the amount of exercise.

2) Assessment: Assessment is an instrument to measure how well students were able to apply the four skills in the context of the constructions and programming templates covered in the instruction. The Assessment measured the participants' ability to read and write semantics and templates. It consisted of seven items that increased in difficulty, based on face validity and the execution of pilot tests with beginning programmers.

The items assessed specific programming skills.

- The items that assess *reading semantics* asked students to track independent code segments that were not part of a larger code base, determine which initial program state would result in a given final state, and comment on their code. Only the experimental group had practice on them;
- The items that assess the *writing semantics* asked students to translate a description of the program's steps and asked them to write the correct syntax; while the two groups practiced these questions, the control group had more practice;
- The items that evaluate *reading templates* asked students to summarize in natural language what a program did. Only the experimental group had practice on them;
- The items that assessed *writing templates* provided students with a description of the problem (ambiguously) and asked them to write a plan in natural language to solve the problem; while the two groups practiced these questions, the control group had more practice.

C. Preparation

There was no need to buy any tools for this experiment. However, it was necessary to print 30 copies of the following materials: pre and post-study survey, Study Material, and Assessment.

D. Data Analysis

During the data organization and processing procedure, we use the RStudio software as a central tool in the data analysis process, given the power of this tool to perform complex statistical calculations, assisting in the processing of data transformation into information.

E. Threat Analysis

We consider some factors that generated threats and directly influenced the conclusions of this paper. Between them:

- We translated the instruments into Brazilian Portuguese, so to minimize possible translation errors, each device was reviewed by two external researchers;
- Problems related to incorrect interpretation of questions;
- Survey participants may be intimidated or uncomfortable performing the tests. We apply the guidelines of the research ethics committee to minimize this possible constraint. The Human Research Ethics Committee of the University Federal University of Campina Grande (UFCG) approved this research (Protocol: 55160816.6.0000.5182). Only participants who signed the consent form participated in the study;
- The instruments were corrected manually, so to mitigate possible human errors, and we double-checked the responses. However, it is possible to think of applications and computerized corrections of this type of Instrument to reduce eventual errors;
- Like all empirical research, this work has threats to validity. The number of subjects participating in the study does not allow generalization of results;

- A considerable sample that allows the formation of a database that, according to a psychometrist, based on probability, statistics, and axioms of the measure and considering the objective of the Instrument, can have centralized control in the application.

F. Research Execution

We conducted the study simultaneously in two classrooms, with participants separated by the condition (control and experimental). We performed a set of steps during the experimental process:

- We explained the purpose of the research to participants who signed a consent form to participate in the study;
- We conducted a pre-survey with questions about their trust in Python;
- During the 3-hour instructional time, we provided participants with study material with instructions, practice problems, and solutions based on different experimental conditions. Students worked on the content at their own pace, and we encourage them to work sequentially through the material and try to solve problems before looking at solutions. As a way to track their progress, we asked participants to write their initials on the page they completed. Participants could take breaks and ask questions, while only questions related to the content of the material we answered, and questions asked about the practice we responded to after the study completed;
- Participants then spent 60 minutes to complete an assessment test to measure how much they learned;
- Finally, students conducted a post-survey with questions about their confidence in Python and provided feedback on the study;
- We transform the answers given during the application into dichotomous items (right/wrong), assigning 0 to make mistakes and 1 to get it right.

V. Results

This section presents the result of our statistical analysis with the collected data.

A. Reliability

We analyzed the reliability of the instruments through the internal consistency (correlation between different items in the same Instrument) and the psychometric properties of items. This procedure is an essential step in the construction of any instrument, as it allows to check if the constructed scale is minimally adequate to continue the study, that is, to verify through this study if the instruments add sufficient reliability if not, it must be improved as to its specificity's.

1) Internal Consistency: We calculated the internal consistency of the instruments using the ltm package with the function *cronbach.alpha()*; the output can be seen in the Tables II and III. From the values obtained for Cronbach's alpha, we conclude that the results obtained in the Study Material instrument are reliable for all the skills presented. In the Assessment, it is necessary to adjust/create new items for reading templates and writing templates.

TABLE II
Study Material Internal Consistency

Skill	Subjects	Items	Average	Standard Deviation	Cronbach's Alpha
Reading Semantics	30	39	34.412	8.872	0.977
Writing Semantics	15	37	32.179	8.999	0.980
Reading Templates	30	24	16.902	6.968	0.952
Writing Templates	15	9	5.564	2.474	0.748

TABLE III
Assessment Internal Consistency

Skill	Subjects	Items	Average	Standard Deviation	Cronbach's Alpha
Reading Semantics	30	10	2.300	2.208	0.796
Writing Semantics	30	4	1.300	1.215	0.710
Reading Templates	30	2	1.033	0.795	0.633
Writing Templates	30	2	0.867	0.763	0.588

2) Psychometric Properties: We interpreted the distribution of participants' responses in each skill through the 2PL [19] [20]. We estimated item parameters, the proportion of correct answers, and the biserial point correlation between the correct answer in the item and the total score in the task.

We calculate the parameters using the ltm package with the *tpm()* function. Using the same package, we calculated the biserial point correlation of the items with the *biserial.cor()* function, a part of the output we present in Table IV, where we present the first five items out of 10 that make up the reading semantics skill assessment with the parameters we consider in this paper. The complete table of items and all study material skills and assessment tools are available at https://github.com/yesjucelio/applying-irt-santos.

TABLE IV
Assessment Reading Semantics Skill Psychometric Properties, first 5 calibrated items

Id	Item	a	b	p.a.	c.p.b
001	1	0.837	-0.589	0.600	0.247
002	2	2.200	-0.260	0.567	0.511
003	3a	2.773	0.599	0.300	0.613
004	3b	1.476	1.766	0.133	0.450
005	3c	3.031	1.559	0.300	0.631
Average		2.063	1.260	0.230	0.512
Standard Deviation		0.880	1.076	0.208	0.128

We present in Figures 1 and 2, respectively, the graphical representation of the ICCs and IIF of the first five calibrated items for the reading semantics Skill task, where we highlight the extreme values of the discrimination and difficulty indices and how much information each item provides in a specific region of the latent trait. We offer full item images and all study and assessment material skills at https://github.com/yesjucelio/applying-irt-santos. We plot the ICCs and IIF through the ltm package and using the *plot()* function [4].

We evaluated the required conditions of the item parameters to 2PL as a way to avoid compromising the representatives of the evaluated domain. We did not find critical values for the estimated parameters, in both instruments, all estimated skill items have values above 0.30 and below 4 for the discrimination index and values between 3.95 and -3.95 for the item difficulty.

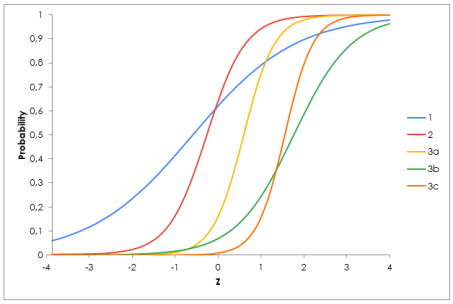

Fig. 1. Assessment Reading Semantics Skill ICC, first 5 calibrated items

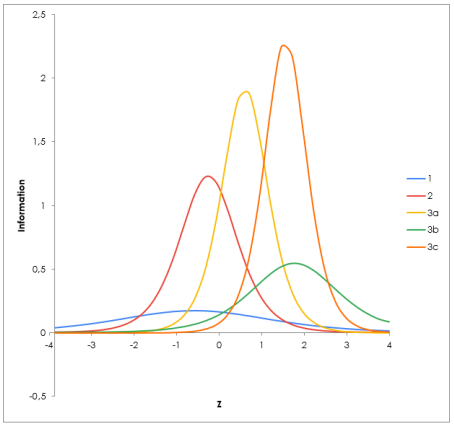

Fig. 2. Assessment Reading Semantics Skill IIF, first 5 calibrated items

The graphical representation of the characteristic curves of the items for the estimated skills in both instruments highlights the extreme values of item discrimination and difficulty indices. The discrimination parameter is proportional to the slope of the tangent line of the item's characteristic curve at point b (item difficulty), the more inclined the curve, the higher the item's discrimination. Regarding the difficulty parameter, an item is difficult when the ICC is positioned on the left, the further away, the greater the difficulty of the item [4]. In the example in Figure 1, item '3c' is the most discriminating item '3b' is the most difficult, and item '1' is the easiest and least discriminatory.

Regarding the IIF, we can see how much information each item provides in a given latent trait region for game activities. In the example in Figure 2, item '3a' gives less information when the skill is at -2, by comparison. However, this is the item that provides more information in the area where the skill is average (when theta is close to 0).

B. Validity

We analyzed the performance of the groups in the reinforcement and evaluation activities. We want to investigate whether explicitly indicating skills and the practice of using each skill improves student performance in reading and writing code in the reinforcement and assessment activity?

To answer this question, we estimated the skills of participants who answered the study material and Assessment in both groups. Figures 3 and 4 provides a summary of the data[2].

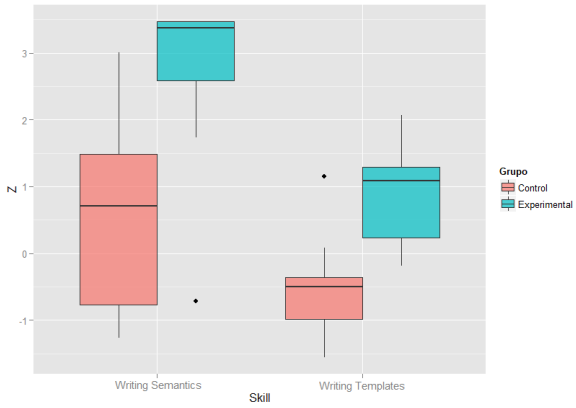

Fig. 3. Writing Semantics and Writing Templates Skill's Boxplot graph, between groups in Study Material

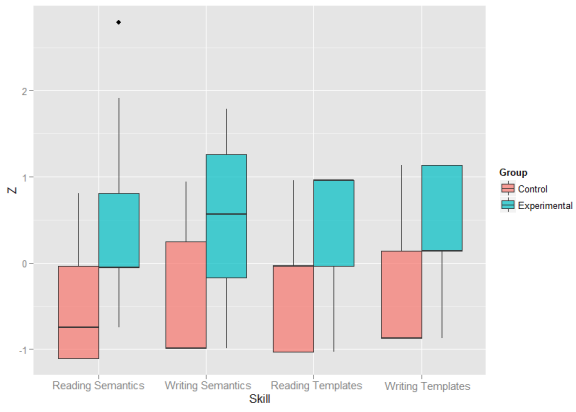

Fig. 4. Skill's Boxplot graph, between groups in Assessment

[2]Due to the Study Material of the control group not working on reading skills, it was not possible to plot the graphics for these skills, only to have the items in stock in both materials

In Figure 3, despite some discrepant values, the experimental group showed less variation when compared to the control group concerning the writing semantics skill. When compared to the other skill, this variation was more significant in the experimental group. Overall, we can see that in writing semantics and writing templates, many individuals in the experimental group had higher skill estimates when compared to the control group.

We observed this behavior in Figure 4, where all the skills of individuals in the experimental group are superior when compared to the control group. Tables V and VI show the means and standard deviation of the performance of both groups in the study material and the Assessment, respectively. The average estimate of participants' skills in both instruments is higher in the experimental group.

TABLE V
STUDY MATERIAL SKILL'S ANALYSIS, BETWEEN GROUPS

Skill	Control Group		Experimental Group		P-value
	Median	Standard Deviation	Median	Standard Deviation	
Writing Semantics	0.710	1.395	3.374	1.118	0.0001
Writing Templates	-0.467	0.670	1.090	0.759	0.0001

TABLE VI
ASSESSMENT SKILL'S ANALYSIS, BETWEEN GROUPS

Skill	Control Group		Experimental Group		P-value
	Median	Standard Deviation	Median	Standard Deviation	
Reading Semantics	-0.743	0.701	-0.049	0.940	0.0010
Writing Semantics	-0.987	0.735	0.567	0.916	0.0142
Reading Templates	-0.036	0.718	0.963	0.733	0.0068
Writing Templates	-0.873	0.746	0.140	0.675	0.0083

Then, we checked the distribution of our dataset. We applied the Shapiro-Wilk test, with a significance level of 5%, to observe whether the data set follows a normal distribution.

In Study Material, we obtained a p-value = 0.0013 for writing semantics and a p-value = 0.01835 for writing templates. In Assessment, we obtained a p-value = 0.01 for reading semantics, a p-value=0.0013 for writing semantics, a p-value = 0.0001 for reading templates and a p-value = 0.0001 for writing templates. Thus, we can conclude that both samples do not follow a normal distribution, with a confidence level of 95%.

Given the non-normal distribution of data, we used the Mann-Whitney to compare the estimated skill of participants for both groups who responded to skills writing semantics and writing templates, in Study Material, according to Table V. And, all skills, in Assessment, according to Table VI.

We consider as null hypothesis that the estimate for each skill of the experimental group is lower when compared to the control group for each of the instruments (Study Material and Assessment). Based on the p-values in Tables V and VI, we can conclude that the performance of participants who practiced the reading semantics and reading templates skills are performed better in writing semantics and writing templates skills when compared to participants in the control group, both in the Study Material and in the Assessment.

VI. DISCUSSION

In this section, we discuss the previous results. Through CTT and IRT, we determined some indications of the validity of the instruments construct. Also, we measure students' abilities by these theories to achieve the most satisfactory results. Thus, we sought some initial evidence to assess the effect of such instructions on these skills and identified the amount of practice required for each skill.

A. Are the Instruments Reliable?

Through CTT, we calculated the biserial point correlation and Cronbach's alpha coefficient to estimate the internal consistency of the instruments.

High biserial point correlation values contributed to the right internal consistency of the instruments. High item-total correlations indicate that items associate closely with each other. As expected, Cronbach's alpha values were acceptable for all skills worked on in the study material. However, in the evaluation, the skills reading templates and writing templates had lower indications than expected, presenting a weak correlation of items and, consequently, little internal consistency. This result may come from a small number of items to measure these constructs. Therefore, it is necessary to construct more items for these skills to improve the internal consistency of the evaluative Instrument.

Using IRT, we estimate the psychometric properties of items. Difficulty parameters have close values regardless of the sample, differing when it is estimated with much error, such as when the sample is composed of subjects with no variation in skill levels. In short, this result also reflects the complexity involved in estimating the difficulty of an item based solely on the tacit knowledge of the designers, without considering the empirical results. We evaluated the required conditions of the item parameters for the 2PL model to avoid domain compromise. We did not find critical values for the estimated parameters; in both instruments, all estimated skill items have values within the established parameters slope and threshold.

B. Do the Instruments have Evidence of Validity?

The results provide some initial evidence to suggest practicing each of the four programming skills. We found that participants who received practice in all four skills obtained a better estimate of study material and Assessment, especially for the most advanced knowledge of programming introduction. These results reinforce the theory created by Xie et

al. [2], where providing explicit, sequential instructions for programming skills helps beginners learn to program when compared to more focused writing instructions that do not distinguish these skills. One explanation for such results is the theory itself that reinforces knowledge incrementally.

By practicing semantic reading, the student tracks or code (isolating him from other skills), this practice involves examining fixed-code items in which the student determines the intermediate and final states of the program to a particular standard [21] [6]. Semantic reading is a precursor skill to write syntax and use templates; therefore, it is the basis for all other skills [22]. When students have a strong understanding of reading semantics for a particular programming construct, students should be able to understand how this construct affects program instruction and the output of a piece of code.

After practicing reading semantics and receiving feedback in the form of the correct solution and an explanation, students learn to write the right syntax. To write correct code that meets an unambiguous specification, the student must have an understanding of reading semantics to know how code constructs affect execution and an understanding of writing semantics to understand how to translate these code constructs to correct syntax [2].

Behind learning how to read and write semantics for a new construct, the student moves on to using standard code usage pattern templates to apply knowledge of that construct. Templates are an abstraction from the programming knowledge that has generality and reuse. A template consisted of an objective as well as the various parts or steps necessary to make the model fulfill its intended purpose [23] [24]. In total, we taught four templates: variable swap, digit processing, float equality, and max/min.

After getting involved with instructions on how to read a model, the lesson began to teach how to use a model to fulfill a computational goal. Troubleshooting is essential for this process, but it is beyond the scope of this statement. While the semantic writing statement specified rules to avoid syntactic errors, the writing model statement specified regulations to prevent logic errors that would be syntactically correct but would result in code that did not meet the model specification [2].

VII. Conclusions and Future Works

In this exploratory and preliminary study, we explored IRT in the instruments used in the experiment to verify the internal consistency and psychometric properties of the items. Study Material instrument is reliable for all presented skills. However, in Assessment, it is necessary to adjust/create new items for skills reading templates and writing templates. The items present in the instruments have excellent psychometric properties capable of measuring well the investigated construct.

Then, we investigate whether explicitly indicating skills and the practice of using each skill improves student performance in reading and writing code in the reinforcement and assessment activity. Our dataset suggests that the performance of participants who practiced the reading semantics and reading templates skills are performed better in writing semantics and writing templates skills when compared to participants in the control group, both in Study Material and Assessment. Moreover, the confidence level in Python increased in both groups, but this increase was not significant.

As future work, we plan to deepen our IRT research to assess introductory programming skills. We will extend the experiment to a more extensive study to address advanced knowledge anticipated in introductory programming disciplines. Develop new items for study and evaluation material, as well as improve the reliability index of these instruments. To have strong evidence of the theory created by [2], we will apply this study to newcomers in the Computer Science course who have never had contact with programming. We hope that the results of this research can contribute to the development and discussion of the Assessment of introductory programming skills, as well as the effort to use IRT in the construction/validation of reliable instruments in Computer Science.

References

[1] S. Wang, Y. Han, W. Wu, and Z. Hu. Modeling student learning outcomes in studying programming language course. In *Proceedings of the International Conference on Information Science and Technology (ICIST)*. IEEE, 2017.

[2] B. Xie, D. Loksa, G. L. Nelson, M. J. Davidson, D. Dong, H. Kwik, A. H. Tan, L. Hwa, M. Li, and A. J. Ko. A theory of instruction for introductory programming skills. *Computer Science Education*, 29(2-3), 2019.

[3] J. C. Nunnally. *Psychometric Theory 3E*. Tata McGraw-Hill Education, 1994.

[4] A. L. S. O. Araújo, J. S. Santos, M. R. A. Melo, W. L. Andrade, D. D. S. Guerreiro, and J. C. A Figueiredo. *Metodologia de Pesquisa em Informática na Educação: Abordagem Quantitativa de Pesquisa*, chapter Teoria de Resposta ao Item. SBC, Porto Alegre, 2019.

[5] L. Pasquali. *Psicometria: Teoria dos Testes na Psicologia e na Educação*. Editora Vozes Limitada, 2017.

[6] R. Lister, T. Clear, D. J. Bouvier, P. Carter, A. Eckerdal, J. Jacková, M. Lopez, R. McCartney, P. Robbins, O. Seppälä, and E. Thompson. Naturally occurring data as research instrument: Analyzing examination responses to study the novice programmer. *ACM SIGCSE Bulletin*, 41(4), 2010.

[7] G. R Bergersen, D. I. Sjøberg, and T. Dybå. Construction and validation of an instrument for measuring programming skill. *IEEE Transactions on Software Engineering*, 40(12), 2014.

[8] S. O. Araújo, A. L, J. S. Santos, W. L. Andrade, D. D. S. Guerrero, and V. Dagiene. Exploring computational thinking assessment in introductory programming courses. In *Proceedings of the Frontiers in Education Conference (FIE)*. IEEE, 2017.

[9] M. Lopez, J. Whalley, P. Robbins, and R. Lister. Relationships between reading, tracing and writing skills in introductory programming. In *Proceedings of the fourth International Workshop on Computing Education Research*, 2008.

[10] R. Lister, C. Fidge, and D. Teague. Further evidence of a relationship between explaining, tracing and writing skills in introductory programming. *Acm SIGCSE Bulletin*, 41(3), 2009.

[11] R. Primi. Psicometria: Fundamentos matemáticos da teoria clássica dos testes. *Avaliação Psicológica*, 11(2), 2012.

[12] F. B. Baker. *The Basics of Item Response Theory*. ERIC, 2001.

[13] J. M. Andrade, J. A. Laros, and V. V. Gouveia. O uso da teoria de resposta ao item em avaliações educacionais: Diretrizes para pesquisadores. *Avaliação Psicológica*, 9(3), 2010.

[14] S. E. Embretson and S. P. Reise. *Item Response Theory*. Psychology Press, 2013.

[15] D. F. Andrade, H. R. Tavares, and R. Cunha Valle. *Teoria da Resposta ao Item: Conceitos e Aplicações*. ABE, São Paulo, 2000.

[16] J. P. Fox and C. A. W. Glas. Bayesian estimation of a multilevel irt model using gibbs sampling. *Psychometrika*, 66(2), 2001.
[17] I. R. R. Lu, D. R. Thomas, and B. D. Zumbo. Embedding irt in structural equation models: A comparison with regression based on irt scores. *Structural Equation Modeling*, 12(2), 2005.
[18] M. J Kolen and Y. Tong. Psychometric properties of irt proficiency estimates. *Educational Measurement: Issues and Practice*, 29(2), 2010.
[19] F. B Baker and S. Kim. *Item Response Theory: Parameter Estimation Techniques*. CRC Press, 2004.
[20] W. J. Van der Linden and R. K. Hambleton. *Handbook of Modern Item Response Theory*. Springer Science Business Media, 2013.
[21] R. McCartney, J. E. Mostrom, K. Sanders, and O. Seppala. Questions, annotations, and institutions: Observations from a study of novice programmers. In *Proceedings of the Conference on Computer Science Education*. Helsinki University of Technology, 2004.
[22] G. L. Nelson, B. Xie, and A. J. Ko. Comprehension first: Evaluating a novel pedagogy and tutoring system for program tracing in cs1. In *Proceedings of the Conference on International Computing Education Research*. ACM, 2017.
[23] M. C. Linn and M. J. Clancy. The case for case studies of programming problems. *Communications of the ACM*, 35(3), 1992.
[24] M. J. Clancy and M. C. Linn. Patterns and pedagogy. *ACM SIGCSE Bulletin*, 31(1), 1999.