Screening Programming's Reliability to Measure Predictive Programming Skills

Danilo Medeiros Dantas[1], **Jucelio Soares dos Santos**[1] **Kézia de Vasconcelos Oliveira Dantas**[1]
Wilkerson L. Andrade[2], **João Brunet**[2], **Monilly Ramos Araujo Melo**[2]

[1]State University of Paraíba (UEPB), Campina Grande, Brazil

[2]Federal University of Campina Grande (UFCG), Campina Grande, Brazil

danilo.dantas@aluno.uepb.edu.br, {jucelio, kezia}@servidor.uepb.edu.br

{wilkerson, joao.arthur}@computacao.ufcg.edu.br, monillyramos@gmail.com

Abstract. *This study aimed to evaluate the reliability of an item bank developed in the Screening Programming system for measuring predictive programming skills. The results revealed that the selected items showed good content analysis and consistent psychometric properties. Furthermore, the instruments created from this item bank demonstrated good reliability in professional assessments, validating their accuracy and stability across different contexts and populations. These findings contribute to the programming field by providing a reliable instrument for assessing and developing predictive skills in this domain, fostering continuous advancements in understanding and teaching these skills.*

1. Introduction

In the previous study [Santos et al. 2022], we categorized predictive programming skills and explored instructional approaches to foster and measure these skills. Predictive skills encompass Problem-Solving, Abstract Thinking, Mathematical Reasoning, and Cognitive Flexibility. The field of programming education employs diverse approaches based on educational theories, instructional frameworks, and methodologies to develop these skills. However, there is a need for more valid technologies and instruments to measure predictive programming skills effectively.

To address this gap, we developed Screening Programming, a web-based system designed to accurately and comprehensively assess predictive programming skills. The system includes an item bank carefully crafted to capture the indicators of these skills. We aim to provide a reliable set of assessment tools that can advance the field and facilitate the development of effective teaching and evaluation strategies. The primary objective of this paper is to analyze the validity of the items in the Screening Programming item bank, ensuring that they adequately encompass the indicators of predictive programming skills.

Additionally, we conducted item calibration to gather information related to the specific psychological construct being measured. This step is crucial to ensure that the selected items effectively assess predictive programming skills. Moreover, we assessed the reliability of the instruments by applying Classical Test Theory (CTT) and Item Response Theory (IRT) [Barker 2010, Araújo et al. 2019]. This comprehensive analysis offers insights into the reliability of the instruments and aids in interpreting the results obtained.

We organized this paper as follows: Section 2 presents related studies on measuring and developing predictive programming skills. Section 3 discusses the methods employed for item analysis, calibration, and reliability evaluation using CTT and IRT. Sections 4 and 5 present the results obtained and their implications in light of the existing literature, contributing to the advancement of predictive programming skills and offering insights for developing more effective teaching and evaluation strategies in this field. Finally, Section 6 concludes the paper and suggests avenues for future research.

2. Related Work

This section reviews related studies that address the measurement and development of predictive programming skills and the strategies used. The literature review will allow us to contextualize this study to existing contributions, identify relevant research gaps, and discuss the limitations encountered in previous work.

Several studies have explored predictive programming skills and their measurements [Attallah et al. 2018, Cabo and Lansiquot 2014, Durak 2018, Jones and Westhuizen 2017, Smetsers and Smetsers 2017]. These studies conducted analyses to identify the key indicators of predictive programming skills, highlighting the importance of Problem-Solving, Abstract Thinking, Mathematical Reasoning, and Cognitive Flexibility. While these findings provide valuable insights, the need for more consensus or conflicting findings among these studies suggests further investigation and clarification of the indicators of predictive programming skills.

Another common approach is using educational theories to teach predictive programming skills. A study investigated different educational theories applied to programming education and their influences on skill development [Skalka and Drlík 2018]. These studies emphasize the importance of adapting teaching strategies to theories that support effective learning. However, the limitations and challenges associated with applying these theories in practice require further exploration and understanding.

Several studies have used CTT to assess predictive programming skills. For example, a study applied CTT to assess the programming skills of a group of students, providing a detailed analysis of the results [Durak 2020]. However, CTT has certain limitations, such as its reliance on item difficulty estimates and assumptions of item independence, which may affect the validity and reliability of the assessments. Alternatively, some researchers have adopted approaches, such as IRT, to measure these skills [Jakoš and Verber 2017]. While IRT offers advantages in modeling individual item characteristics, it also has limitations we must consider to gain a comprehensive understanding of assessment methods.

While there is interest in measuring predictive programming skills, there are challenges regarding the availability of valid technologies and instruments. The Screening Programming web system in this study contributes significantly to this context. However, we present the specific limitations or challenges we encountered during this system's development and its bank of items.

In summary, the related work addresses different aspects of predictive programming skills, from identifying indicators to measuring and fostering these skills. However, the lack of consensus among studies, the limitations of assessment methods, and the challenges in developing valid technologies and instruments are areas that require further investigation. The review of related work highlights the relevance of this research and indicates possible directions for future studies.

3. Methodology

In this section, we describe the methodology for evaluating the reliability of the Screening Programming item bank.

3.1. Item Bank Construction

This section presents the planning of the item bank construction. We will address how we answered the following research question:

- **RQ1.** Do the items encompassing predictive programming skill indicators have good content and semantic analysis?

We employ psychometric theories to answer this research question [Araújo et al. 2019, Baker and Kim 2017]. These theories provide a set of steps we explore to construct valid items, as specified below.

Together with a multidisciplinary team of experts, we developed an item bank for the instruments. We designed the items to cover easy, moderate, and difficult Problem-Solving levels, thus divided into three levels. We began by drafting the items, describing the alternatives (in the case of multiple-choice items), and indicating the correct response, ensuring that the examinee can express their skill by answering the item.

Item development also involves determining the type and quantity of items composing the instrument. The type of item depends on the instrument's purpose. Therefore, in developing the items, our team considered the number of items managed by the instrument (these details depend exclusively on the item type). Thus, the larger the items number in the bank, the better the instrument, as there will be more suitable items for a particular skill level. Overall, we constructed 40 items to compose the assessment of predictive programming skills.

After item construction, we conducted theoretical evaluations. Judges performed theoretical analyses as experts in the research field. The judges examined whether the items were well understood (semantic analysis) and suitable for measuring the desired skill (content analysis).

Semantic analysis verifies whether the items are intelligible to all individuals. The items must be easily understandable for everyone, even those with lower skill levels. The semantic analysis aims to ensure that the difficulty in understanding the items is not a complicating factor that may interfere with the subject's response.

The content analysis aims to ensure that the items are related to the skill we intend to estimate. The judges' number may vary, but the literature recommends having at least five [Araújo et al. 2019, Baker and Kim 2017]. Therefore, we used five judges to analyze the questions' content. The literature also indicates that an agreement 80% among the judges can serve as a reference for deciding whether to include the item in the instrument. If the agreement is below 80%, the item should be excluded from the instrument. In the case of five judges, at least four include the item in the instrument.

3.2. Item Bank Calibration

This section presents the planning of item bank calibration. We will address how we answered the following research questions:

- **RQ2.** Do the items enclosing predictive programming skill indicators have good psychometric properties?
- **RQ3.** Do the instruments developed from the item bank demonstrate good reliability?

After item preparation, we proceeded to the item bank calibration stage. In this stage, the items had already undergone semantic analysis and content analysis by the judges. Now, these items would proceed to an item bank.

Calibrating the item bank involves applying the items, collecting data, selecting the response model, using the calibration method, designing the scale, and interpreting the scale. When applying the items, we needed a sufficient sample of respondents. The sample size depends on the item number in the bank, so the more significant the item number, the larger the sample size. For a database of 10 items per skill, we needed a sample of at least 100 respondents, i.e., at least ten times for each item. We could have used the traditional paper-and-pencil format, but we presented it in a computerized version to obtain the sample. All subjects were required to respond to all items in this stage.

Thus, we applied the items to 100 students from local community colleges in Computer Science. We explained the study to teachers and administrators and requested a sample of students who fit the study's profile. We considered this study's inclusion criteria: Assent Form - if the student is under 18 years old - or signed Informed Consent Form; and absence of cognitive, visual, psychological, or neurological impairments. We excluded individuals who did not sign the authorization term from the study and had cognitive, visual, psychological, or neurological impairments.

We considered the students' responses during the instrument application and transformed them into dichotomous (0/1) during scoring, assigning 0 for incorrect and 1 for correct responses. We then tabulated the data and conducted a study to verify the psychometric properties of the items. This procedure is essential in constructing any instrument as it allows us to determine if the constructed scale is minimally suitable for further study.

The choice of the response model involves determining which IRT model fits the instrument. We checked if the fit was appropriate and, if necessary, replaced the misfitting model. A poorly fitting model will not provide consistent parameters for items and abilities. If the item parameter estimates through IRT are inconsistent, for example, exhibiting outlier values or high standard error, it may be due to inadequate sample size.

The item bank calibration method estimates item parameters using CTT and IRT criteria. Analyses using CTT indices helped eliminate inadequate items. We performed the item calibration process using IRT and parameter analysis. This analysis assisted in the decision to exclude inadequate items. We estimated the parameters using a Bayesian approach. After eliminating inadequate items, we performed the analyses again using CTT and dimensionality indicators to verify that they were not affected by the exclusion of items. Next, we re-executed the calibration process to check if the remaining items were appropriate and were not affected by the exclusion of other items. IRT considers an item bank well-calibrated if the item parameter estimates are appropriate and the standard errors are low.

Estimates with critical parameters from the 3-parameter logistic model (3ML) imply that we should remove the item from the bank. The slope index below 0.30 is inadequate for an item to have the power to differentiate subjects with different ability estimates. The threshold index below -2.95 or above 2.95 needs to be improved since the skill scale ranges from -3 to 3 in practice. Lastly, the asymptote above 0.40 is considered a critical value. In all these cases, the literature recommends excluding the item from the item bank to preserve the accuracy of the ability estimate.

After the final calibration of the items, we should verify the item bank. We checked if the item number in the item bank was sufficient for instrument administration. According to the instrument's objective, we verify if the items cover the entire content, are well-distributed, and provide adequate information across all ranges of the evaluated latent trait (easy, moderate, and complex items). After calibrating and equating the items, we construct the ability scale. The scale aims to provide a qualitative interpretation of the values obtained by applying the IRT model, thus allowing for the pedagogical interpretation of the competency values.

3.3. Limitations and Mitigation Measures

Our research, although methodologically rigorous, encounters some significant limitations. The sample of 100 students, while suitable for our objectives, may be considered moderate, limiting the generalizability of the results. The exclusion of items during calibration can impact the diversity of the item bank, and the choice of the item response model is crucial. Furthermore, variations in participants' skills can influence the estimates of item properties, particularly in extreme groups.

We have implemented several strategies to mitigate these limitations, including careful participant selection, transparency in item exclusion, rigorous validation of IRT models, and an ongoing commitment to item bank improvement. We acknowledge the importance of external validity and encourage future research to explore the relationships between scores in this item bank and other relevant measures of predictive programming skills.

Additionally, despite contrary instructions, we addressed an additional concern related to the potential for participants to seek online assistance or consult with third parties during the test. To address this issue, we reinforced the guidelines that participants should respond independently and implemented an online monitoring system to identify potential violations of instructions. These measures aimed to ensure the integrity of the collected data and maintain the results' reliability.

4. Item Bank Construction

This section provides a comprehensive overview and analysis of the results obtained during the item bank construction phase, explicitly focusing on evaluating the quality of content analysis within the items. In addition, we address **RQ1** within this section, thereby providing a detailed examination of the findings and their implications.

Together with a multidisciplinary team of experts, we designed easy, moderate, and difficult items to solve. We drafted the items, described the alternatives, and indicated the correct answer. Next, we analyzed the content of the items to ensure that they referred to the desired skill.

In this stage, five local university professors and programming education experts participated in the study. The professors evaluated whether the items in the instrument measured the construct of the examinees. To assess the degree of agreement among the judges, we used the agree function available in the irr package in the R language. This agreement reached 90%. However, is it reliable? In order to assess the degree of reliability among the judges, we applied the Fleiss' Kappa statistical test, with an 80% significance level (substantial agreement). We calculated it using the kappam.fleiss function in the irr package in the R language. Table 1 presents the reliability level in classifying items among the skills. The level of reliability among the judges is 0.947 (almost perfect).

Table 1. Reliability Analysis among Judges.

Skill	**Judges**	**Items**	**Kappa**	**Z**	$p - value$
Problem-Solving	5	10	0.899	17.971	0.0001
Abstract Thinking	5	10	0.945	18.904	0.0001
Mathematical Reasoning	5	10	1.000	20.000	0.0001
Cognitive Flexibility	5	10	0.945	18.904	0.0002
Reliability	**5**	**40**	**0.947**	**32.800**	**0.0001**

Based on our analysis, there is a high level of reliability and agreement among the judges. We evaluated their responses and found that all items achieved agreement rates exceeding 80%. Consequently, we have selected these items to construct the instrument, which focuses on assessing Problem-Solving skills, Abstract Thinking, Mathematical Reasoning, and Cognitive Flexibility. These criteria ensure that the instrument comprises reliable and valid measures to capture the targeted constructs effectively.

In order to provide a more comprehensive understanding of the skills assessed in this study, we will present a detailed summary of the items and questions that comprise the four essential skills:

- **Problem-Solving:** We assess this skill through practical situations and challenges requiring students to solve complex problems effectively. This skill includes the analysis of scenarios, the identification of obstacles, and the proposal of strategies to overcome them. Items in this category encompass a variety of domains, ranging from logical problems to real-world challenges;
- **problem Solving:** We assess this skill through practical situations and challenges that require students to solve complex problems effectively. This skill includes analyzing scenarios, identifying obstacles, and proposing strategies to overcome them. Items in this category cover a variety of domains, from logical problems to real-world challenges;
- **Abstract Thinking:** We assess this skill through items that challenge students to break complex problems into smaller parts and solve them independently. Examples include analyzing patterns in numerical sequences or identifying logical relationships within data sets. Students are encouraged to think abstractly, identifying and applying underlying concepts to diverse situations;
- **Mathematical Reasoning:** We assess this skill through items that apply mathematical concepts to solve various problems. Examples include items on numerical calculations, algebra, geometry, and statistics. We designed the items to measure

mathematical knowledge and the ability to apply it in practice to solve everyday problems;

- **Cognitive Flexibility:** We assess this skill through problem items that present solutions from different perspectives and approaches to the same problem. Items in this skill challenge students to think creatively, explore multiple solutions, and adapt to different contexts. This skill includes questions that require the generation of creative alternatives and the evaluation of different strategies.

All items related to these skills are available in Portuguese here.

5. Item Bank Calibration

This section will provide an extensive overview and analysis of the results obtained during the item calibration phase. The main objective of this phase is to assess the psychometric properties of the instrument's items and determine whether the constructed scale is adequately adjusted to proceed with the study. We will specifically address **RQ2** and **RQ3** in this context. The findings and insights gained from this analysis will contribute significantly to the subsequent stages of our research.

We presented the items to the participants for information such as item correctness and response time in the activities. For this study, 100 students from local universities and institutes in Campina Grande, Paraíba, Brazil, participated. We applied the instrument virtually, and the students had 24 hours to respond to the items. We emphasize that participants should respond without online consultation or assistance from course monitors. We considered the students' responses during the administration. We transformed them into dichotomous items (correct/incorrect), assigning a value of 0 for incorrect answers and 1 for correct answers during the scoring process.

The data collected in this phase were analyzed using the IRT with the assistance of the Excel tool available at http://psychometricon.net/libirt/ for analysis and adjustment of the 3ML model through maximum likelihood estimation. The analysis aimed to i) assess the internal consistency of the instrument (presented in Table 2, where Cronbach's alpha values indicate that the instrument is reliable for all presented activities) and ii) estimate the item parameters for the instrument tasks.

Table 2. Instrument Internal Consistency.

Skill	Subjects	Items	Average Score	Standard deviation	Cronbach Alpha
Problem-Solving	100	10	7.174	2.648	0.854
Abstract Thinking	100	10	6.652	2.547	0.786
Mathematical Reasoning	100	10	6.522	2.534	0.757
Cognitive Flexibility	100	10	6.261	2.594	0.791

We interpreted the distribution of students' responses to each item of the tasks using the 3ML model. Additionally, we considered the proportion of correct responses and the point-biserial correlation between the correct item response and the total score on the task. Table 3 presents the items that compose the Problem-Solving skill and their respective parameters. The remaining items for the Abstract Thinking, Mathematical Reasoning, and Cognitive Flexibility skills are available here.

Table 3. Problem-Solving Skill - All Calibrated Items.

Id	Slope	Threshold	Asymptote	Hit Ratio	Biserial Point
Res-001	3.719	-0.885	0.143	0.783	0.771
Res-002	3.367	-0.745	0.169	0.783	0.670
Res-003	1.904	-0.367	0.145	0.652	0.637
Res-004	0.815	-1.926	0.165	0.826	0.340
Res-005	2.014	0.633	0.127	0.391	0.486
Res-006	2.015	0.996	0.123	0.304	0.426
Res-007	1.478	-0.817	0.158	0.739	0.522
Res-008	2.568	-1.805	0.160	0.913	0.595
Res-009	2.568	-1.805	0.160	0.913	0.595
Res-010	2.689	-1.371	0.165	0.870	0.694

After verifying the 3ML logistic model, we did not find critical values for the estimated parameters. All items of the Problem-Solving skill, Abstract Thinking, Mathematical Reasoning, and Cognitive Flexibility have values greater than 0.30 for the slope index; for the asymptote index, values between 3.95 and -3.95, and guessing below 0.40. In addition, the results revealed that all tasks have easy items (with rates above 75%), moderate items (with rates between 50 to 75%), and complex items (with rates below 50%).

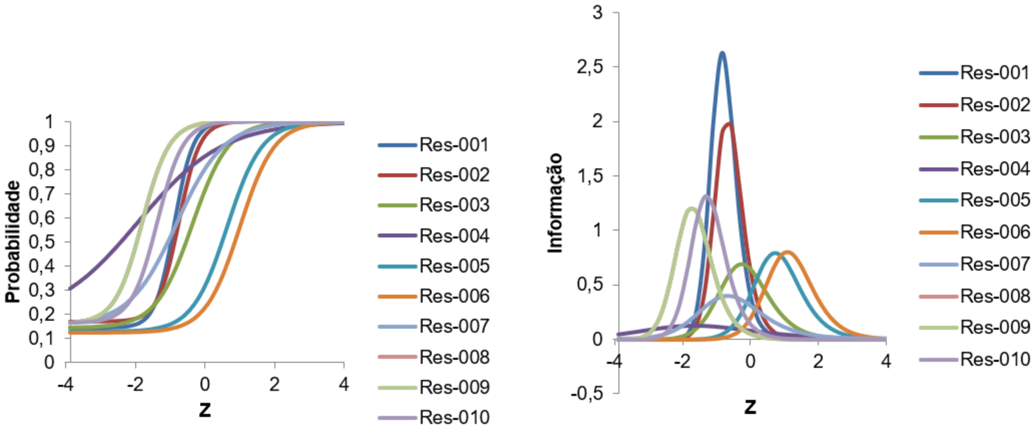
The point-biserial correlations revealed a tendency towards choosing the wrong option by the participants who obtained higher scores in the test for items Res-004, Res-005, and Res-006 in the Problem-Solving skill, for example. Despite this, all items adequately fit the 3ML, thus presenting good reliability and skill separation index. We can apply this same analysis to other skills.

In Figure 1(a), we present the graphical representation of the ICCs of the Problem-Solving skill, highlighting the extreme values of the slope, asymptote, and guessing parameters. Moreover, in Figure 1(b), we present the graphical representation of the FIIs of the Problem-Solving skill, highlighting the amount of information each item provides in a specific region of the latent trait. We provide the ICCs and FIIs of the other skills here.

By analyzing the ICCs, we can observe that item Res-004 is the least discriminative and the easiest to answer. On the other hand, item Res-001 is the most discriminative. The most challenging item to answer is Res-006, and interestingly, it also has the lowest chance of being answered correctly by chance. Item Res-002, on the other hand, is the most accessible item to answer correctly by chance.

Regarding the probability of answering correctly by chance, this skill presents indices below 16%, which is significantly lower than expected, considering that the items for this skill have four alternatives each. Mathematically speaking, the probability of a student with low ability answering the item correctly is approximately 25%.

Figure 1. ICCs and IIFs of Problem-Solving Skill items.



(a) ICCs of all items.

(b) IIFs of all items.

Regarding IIFs, item Res-002 provides more information for evaluating moderate-ability subjects. Therefore, in a computerized instrument, this item would be chosen first if the instrument situates the Theta ability at 0, representing the average ability.

6. Final Considerations

In this study, we aimed to evaluate the reliability of the item bank developed in Screening Programming to measure programming predictive skills. Next, we present the final considerations based on our research questions:

- (**RQ1.**) We developed items with good content analysis that encompass programming predictive skills. Through a systematic literature review and rigorous selection criteria, we created a comprehensive set of items that effectively address the essential indicators of these skills;
- (**RQ2.**) The selected items demonstrated good psychometric properties, highlighting their effectiveness as assessment instruments. During the calibration of the items, we obtained consistent and reliable results, validating their ability to measure programming predictive skills accurately;
- (**RQ3.**) The instruments created from the item bank showed good reliability in professional evaluations. Both the application of the CTT and the IRT yielded reliable results, indicating that the instruments are consistent and stable across different contexts and populations.

In summary, this study significantly contributed to the programming field by providing a validated and reliable item bank to measure skills in this area. We hope this research will foster future investigations and advancements in the field, fostering continuous progress in understanding and developing programming predictive skills.

References

[Araújo et al. 2019] Araújo, A. L. S. O., Santos, J. S., Melo, M. R. A., Andrade, W. L., Guerreiro, D. D. S., and Figueiredo, J. C. A. (2019). *In: Jacques, P. A. and Pimentel, M. and Siqueira; S. and Bittencourt, Ig. Metodologia de Pesquisa em Informática na Educação: Abordagem Quantitativa de Pesquisa*, chapter Teoria de Resposta ao Item. SBC, Porto Alegre.

[Attallah et al. 2018] Attallah, B., Ilagure, Z., and Chang, Y. K. (2018). The impact of competencies in mathematics and beyond on learning computer programming in higher education. In *Proceedings of the Information Technology Trends (ITT)*. IEEE, Dubai, United Arab Emirates.

[Baker and Kim 2017] Baker, F. B. and Kim, S.-H. (2017). *The basics of item response theory using R*. Springer.

[Barker 2010] Barker, T. (2010). An automated feedback system based on adaptive testing: Extending the model. *International Journal of Emerging Technologies in Learning (iJET)*, 5(2).

[Cabo and Lansiquot 2014] Cabo, C. and Lansiquot, R. D. (2014). Synergies between writing stories and writing programs in problem-solving courses. In *Proceedings of the Frontiers in Education Conference (FIE)*. IEEE, Madrid, Spain.

[Durak 2018] Durak, H. Y. (2018). The effects of using different tools in programming teaching of secondary school students on engagement, computational thinking and reflective thinking skills for problem solving. *Technology, Knowledge and Learning*.

[Durak 2020] Durak, H. Y. (2020). Modeling different variables in learning basic concepts of programming in flipped classrooms. *Journal of Educational Computing Research*, 58(1).

[Jakoš and Verber 2017] Jakoš, F. and Verber, D. (2017). Learning basic programming skills with educational games: A case of primary schools in slovenia. *Journal of Educational Computing Research*, 55(5).

[Jones and Westhuizen 2017] Jones, G. B. and Westhuizen, D. V. (2017). Pre-entry attributes are thought to influence the performance of students in computer programming. In *Proceedings of the Southern African Computer Lecturers' Association*. Springer, Cham.

[Santos et al. 2022] Santos, J. S., Andrade, W. L., Brunet, J., and Melo, M. R. A. (2022). A systematic literature review on predictive cognitive skills in novice programming. In *2022 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE.

[Skalka and Drlík 2018] Skalka, J. and Drlík, M. (2018). Educational model for improving programming skills based on conceptual microlearning framework. In *Proceedings of the International Conference on Interactive Collaborative Learning (ICL)*. Springer, Cham.

[Smetsers and Smetsers 2017] Smetsers, R. W. and Smetsers, S. (2017). Problem solving and algorithmic development with flowcharts. In *Proceedings of the Workshop in Primary and Secondary Computing Education (WiPSCE)*. ACM Nijmegen, Netherlands.