修士論文

プログラミング初学者の学習可能性を把握するためのツール開発と評価に関する研究

令和 5 年度卒業

三重大学大学院工学研究科
博士前期課程　電気電子工学専攻

学籍番号　422M203
荒木　諒

目次

1章　はじめに

1.1　背景

現代社会の急速なデジタル化とテクノロジーの進化に伴い，プログラミング教育の必要性が高まっている．デジタル技術の普及や職業の多様化により，プログラミングスキルはエンジニアだけでなく，様々な分野で不可欠なものとなっている．プログラミングを学ぶことで，論理的思考や問題解決能力が向上し，将来の職業市場で競争力を維持するための基盤が築かれる．また，創造性を発揮する手段としてもプログラミングが活用され，デジタル時代における基本的なデジタルリテラシーの向上にも寄与する．このような理由から，プログラミング教育は教育体系において不可欠なスキルの一環となり，重要性を増している．しかし，現代のプログラミング教育には，一般的な課程への統合不足，教育のアクセス不均等，資材や技術の不足，統一されていないカリキュラム，教育者の不足などの多様な問題がある．

図 1-1 は三重大学工学部電気電子工学コースでプログラミング演習の授業を担当している先生にアンケートを行い，おおよその学習者のレベル分けを示したものである[1]．プログラミング教育においては，学習者の素養に大きな差があり，素養のない学習者が授業に追いつけず，プログラム作成においても時間の差が生じる．この状況下で，講師は図 1-1 の③に分類される講師による指導を必要としている学習者に時間を費やすことが求められる．また，①に属する演習を早く終えた学習者は，講師からのスキルアップの指導を受け損ね，時間を余らせる問題も生じている．従って，プログラミング教育では各学習者のレベルに合わせた教育アプローチが重要である．

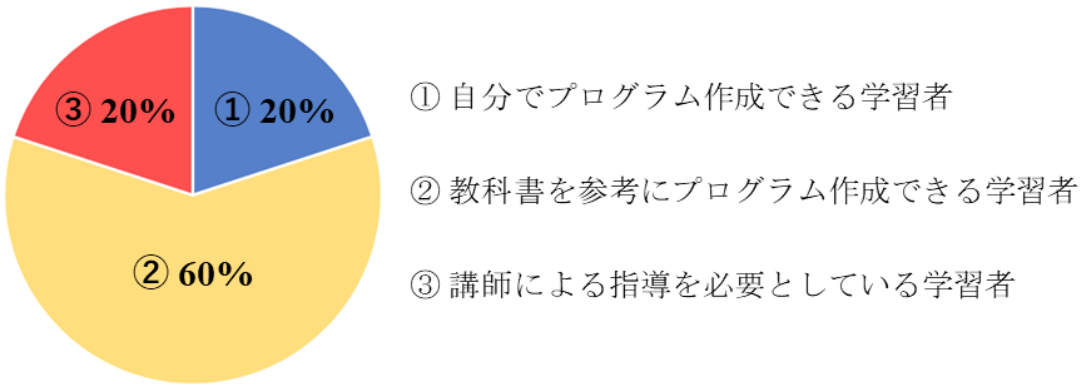


図 1-1 プログラミング学習者のレベル分け

1.2 先行研究

本研究室の小林（2014 年度）は，プログラミングで用いる概念を抽象化した問題によってプログラミング学習前に素養の有無を測るためのプログラミング素養診断テストを考案した[2]．このテストを用いてプログラミング未学習者に適切な教育を行うことが目的である．プログラミング素養診断テストを実施した結果，プログラミングの素養を測ることが期待できるという結論に至ったが，検証実験の対象者が本来の用途として想定している対象とは異なり全員プログラミング経験者であった．このため適切な結果が得られなかった可能性がある．そこで，本研究室の寺久保・大津（2016 年度・2017 年度）はプログラミング初学者を対象にプログラミング素養診断テストを実施した結果，学習者全体の相関は得られなかったが，プログラミング素養診断テストの得点が低い学習者のみを対象とすると相関があり，有用性があることを確認した[3][4]．次に低得点者の相関はあるが高得点者の相関がなかったため，弁別度向上のために素養診断テストに 2 種類の診断項目を追加した（2022 年度）．プログラミング初学者を対象にプログラミング素養診断テストを実施した結果，従来の素養診断テストでは診断されていない素養があることが分かり，診断項目の追加によって，素養診断テストの弁別度向上に期待できることが分かった．なお，素養診断テストで出題する問題や結果などは 2 章で述べる．

1.3 研究の目的

プログラミングは，学習者の素養によって，プログラムの理解度や演習の進捗に大きな差がうまれることが分かっている．そのため，素養診断テストによって，プログラミング初学者の学習可能性を把握し，それぞれの学習者に適した教育を実施することを目的としている．本研究では素養診断テストの学習者の素養の有無を診断する精度を向上させるために従来の素養診断テストのテスト形式を変えたコンピュータ適応型の素養診断テストを提案する．この素養診断テストをプログラミング初学者に実施後，得られたデータの正当性を確認し，実際のプログラミング教育に得られたデータを反映し活用していく．例としては学習者のクラス分けや座席を決める目安としての使用が考えられる．

本論文の構成は以下のとおりである．2 章では従来のプログラミング素養診断テストの問題内容，結果，問題点を述べる．3 章ではプログラミング素養を測る手法の提案，4 章で検証実験の概要について述べ，5 章でその検証実験と結果について述べる．6 章では本手法の活用案について提案する．7 章で本手法の課題について述べ，8 章で本論文のまとめを述べる．

2章 プログラミング素養診断テスト

2.1 目的

プログラミングは学習者の素養によって理解度に大きな差が生まれることが分かっている[1]．そのため，本研究室の小林はプログラミングで用いる概念を抽象化したプログラミング素養診断テストを考案した．プログラミング素養診断テストによってプログラミング初学者の素養を診断し，事前に学習者を識別することができる手法を確立する事を目的とする．本研究では識別不可である素養診断テストの高得点者に焦点を当てる．

2.2 従来のプログラミング素養診断テスト

プログラミング素養診断テストではプログラミングを習得するうえで学習者に必要とされる「新しく学んだ構文や意味を理解して，それを正しく使用する能力」と「間違いを見つけ修正する能力」についての診断をしている．この能力を診断するために素養診断テストは学習者がつまずきやすく，プログラミングの基本である「代入とシーケンス実行」「分岐繰り返し実行」「間違い探し」に関する問題を出題する．ここで，従来の素養診断テストの実施内容を示す．すべて一問一答形式である．

- 代入とシーケンス実行問題　問題数　12 題，　制限時間 12 分
- 分岐繰り返し実行問題　問題数　10 題，　制限時間 8 分
- 間違い探し問題　問題数　10 題，　制限時間 6 分

2.2.1　代入とシーケンス実行問題

代入とシーケンス実行問題では，図 2-1 に示すように箱に数値を格納する命令と 2 つの箱に格納されている数値を入れ替える命令やプログラミングの if のような命令を定義し，図 2-2 に示すような問題を出題している．代入とシーケンス実行問題は処理が上から順番に 1 つずつ実行される点がシーケンス実行を表現している．

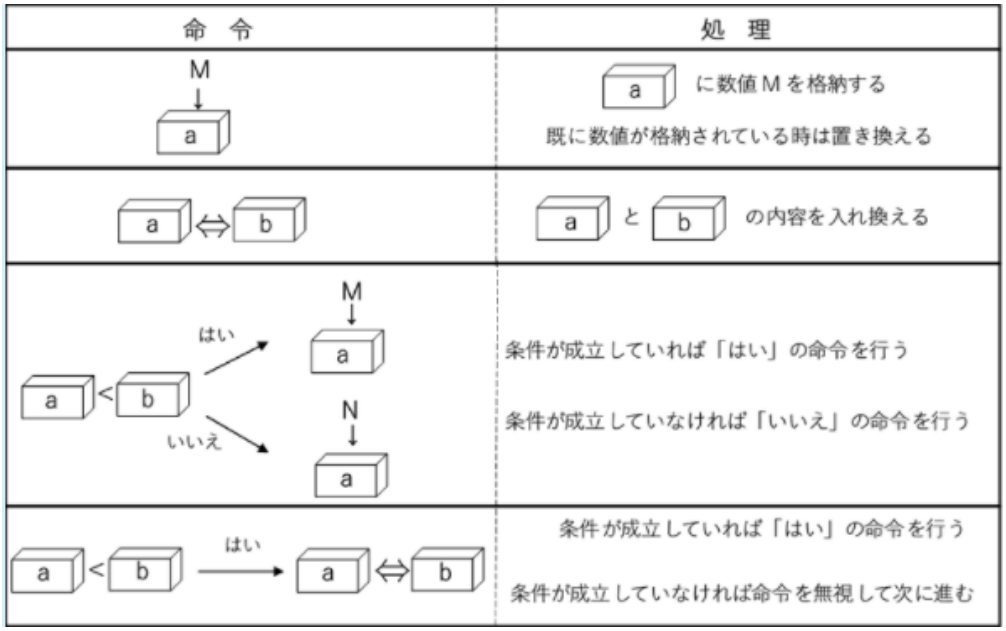


図 2-1 命令の定義

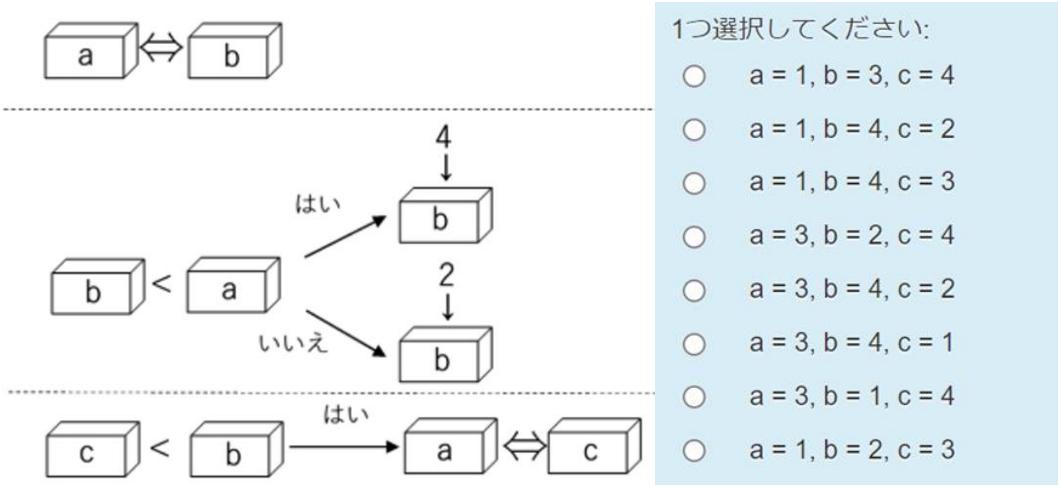


図 2-2　代入とシーケンス実行問題

2.2.2　分岐繰り返し実行問題

分岐や繰り返し実行の素養の有無を測るテストとして，図 2-3 に示すような命令を図 2-3 の中で分岐実行として示される順序に従って処理する問題を出題する．分岐実行・繰り返し実行・分岐繰り返し実行を図 2-4～2-6 に示す．このテストは分岐実行の命令の直前の演算結果が偶数か奇数かによって次の命令が分かれるという処理でプログラミングにおける if 文に近い処理を表現している．これを正しく理解できればプログラミングの分岐処理も理解することができると考えられる．

以上が，小林と大津が考案したプログラミング素養診断テストの概要である．

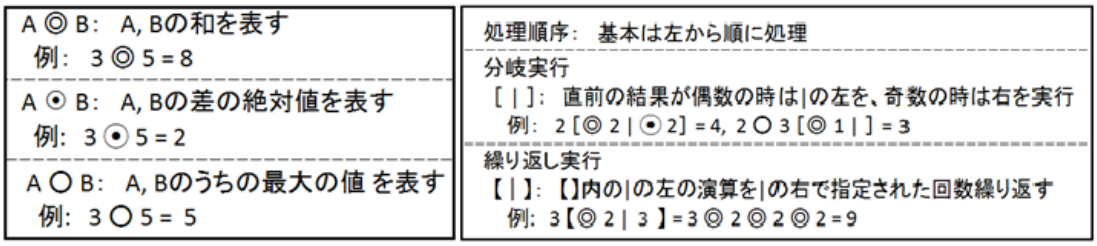


図 2-3　命令の定義

上記の命令を用いると以下の演算結果がどうなるか答えよ

4 ◎ 1[○ 7|◎ 3] = (?)

2 [◎ 2|⊙ 1] [○ 3|⊙ 2] = (?)

図 2-4　分岐実行の問題例

上記の命令を用いると以下の演算結果がどうなるか答えよ

2 ◎ 4【⊙ 10| 2】 = (?)

7 【⊙ 2 ○ 4| 3】 = (?)

図 2-5　繰り返し実行の問題例

上記の命令を用いると以下の演算結果がどうなるか答えよ

7【［◎1 | ◉3］| 3】= (?)

［◉5 | ◎4］【◉2 ○4 | 3】= (?)

図 2-6　分岐・繰り返し実行の問題例

2.2.3　間違い探し問題

プログラミング作成でスペルミスなど簡単なコンパイルエラーを修正できず，プログラムの作成が止まり学習者によって演習の進み具合に差が生まれてしまう．プログラムの作成を円滑に進めるためには間違いを見つける能力が必要である．この素養の有無を測る手法として，図 2-7 に示すような問題を出題する．この問題例は文字列に規則が 5 つ設けられており，その規則に違反していないかを判断する問題である．また， 5 題ごとに規則が変わるため，正確な規則に従う能力が必要とされる．

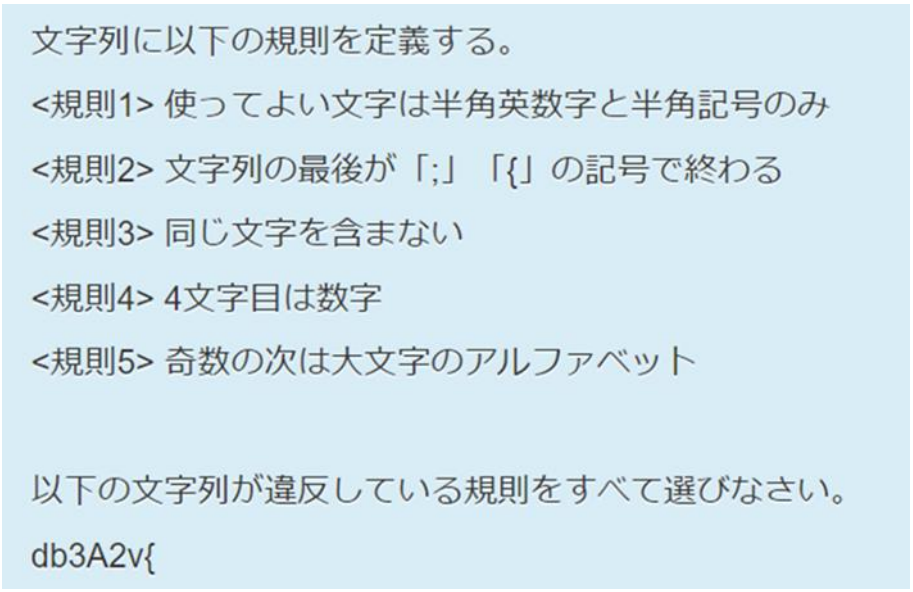


図 2-7　間違い探し問題例

2.3　従来の素養診断テストの結果と問題点

小林・寺久保がプログラミング素養診断テストの検証実験を三重大学の学習者に対して実施している．図 2-8 は 2015 年度に実施した素養診断テストの結果例でプログラミング演習の成績と従来の素養診断テストの相関関係を示している[2]．この結果の相関関係は 0.4 を超える高い相関係数を得られ，相関があることを確認できている．素養診断テストの点数が低い学習者はプログラミング演習の成績も低い結果を得られている．しかし，素養診断テストの点数が高くても，プログラミング演習の成績が低い学習者がいる問題がある．この問題点の考えられる要因として診断できていない素養があることや出題形式に問題があることがあげられる．また，プログラミングの素養はあるがプログラミングにやる気がない学習者がいることも考えられる．

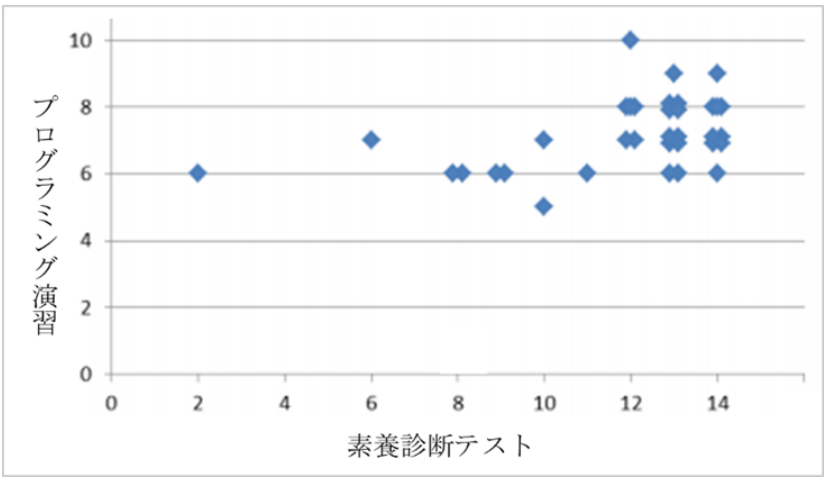


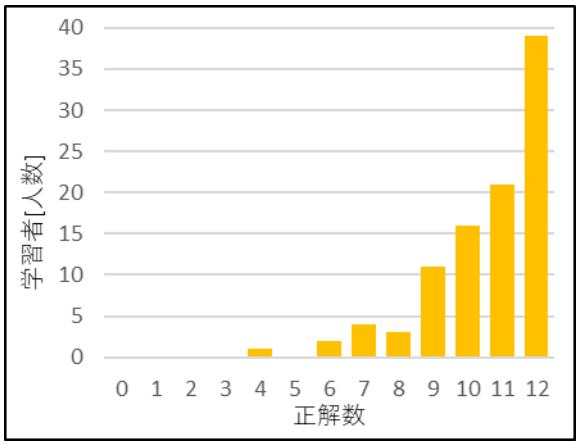
図 2-8 従来のプログラミング素養診断テストの結果

次に，2021・2022 年度のプログラミング言語 I で従来の素養診断テストを三重大学の学習者に実施した結果について，表 2-1 は 2021 年度のプログラミング言語 I の成績，授業点，実際のプログラミング能力の観点から，学習者の事前判定ができているかを確認した結果である．間違い探しの問題追加によって，検出できていなかった要指導学習者を検出していることが分かる．また，3 種類のテストの事前判定で低得点者だった学生は要指導学習者であった．診断項目の追加によってプログラミング素養診断テストの弁別度と精度が向上することを確認することができた．

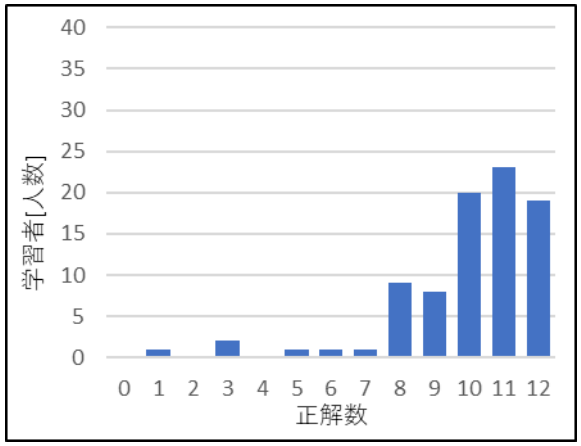
表 2-1 素養診断テスト事前判定結果

	合計	間違い探し	代入シーケンス実行 分岐繰り返し実行	重複
低得点者 （事前判定）	30	16	18	4
要指導学習者 （事後判定）	11	8	7	4

また，図 2-9～2-11 は素養診断テストごとの学習者の得点分布を示す．問題の種類ごとに分布が似ており，すべての分布が高得点に偏っていることが分かる．現状の素養診断テストの問題点は低得点者の識別は出来ているが高得点者の人数が多く識別不可となってしまっている点である．

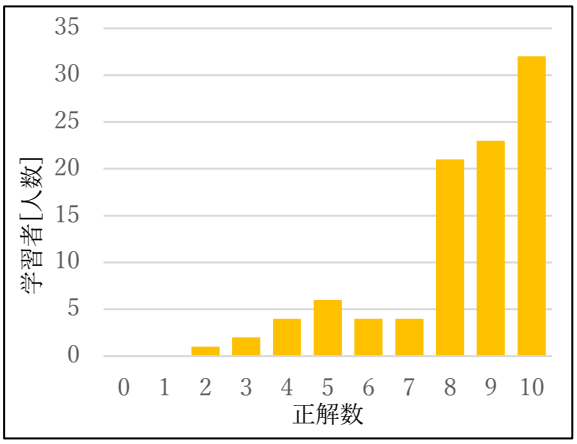


(a) 2021 年度

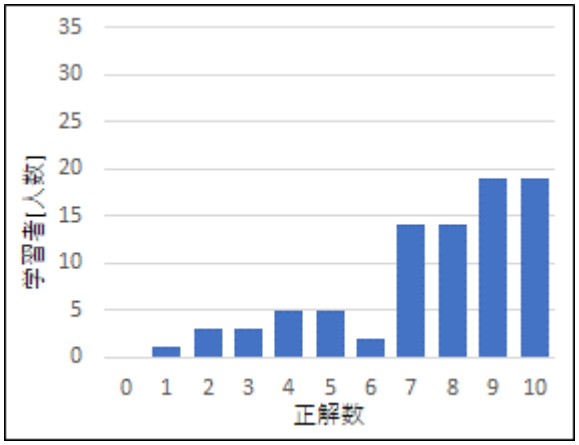


(b) 2022 年度

図 2-9 代入とシーケンス実行問題の結果

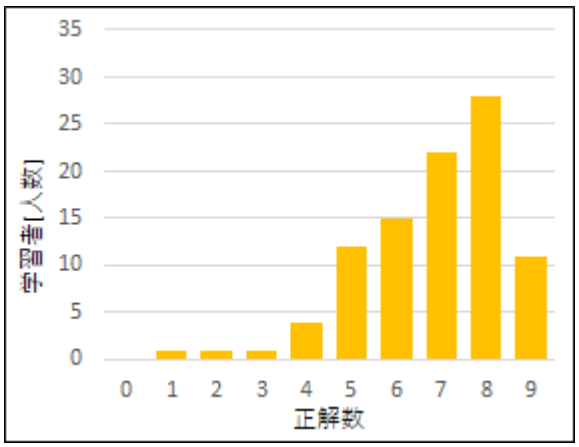


(a) 2021 年度

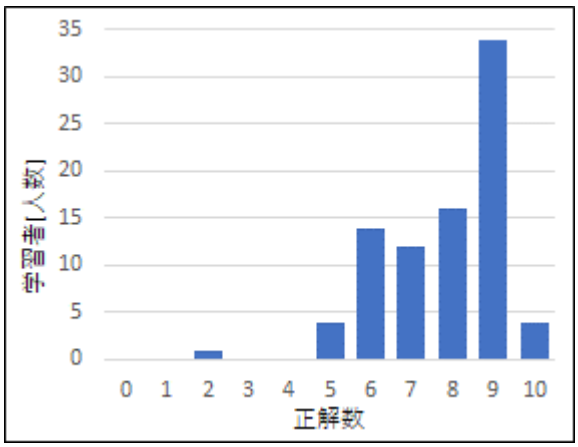


(b) 2022 年度

図 2-10 分岐繰り返し実行問題の結果



(a) 2021 年度



(b) 2022 年度

図 2-11 間違い探し問題の結果

2.4　従来の素養診断テストの課題

従来の素養診断テストは高得点に分布が偏っており，要指導学習者である低得点者の識別は可能だが，演習に時間を余らせてしまう高得点者の識別ができない．また，診断項目を追加することで素養診断テストの精度は向上していくが，素養診断テストは講義時間内に実施するため，講師と学習者に負担をかけないテストを理想としている． 従来の素養診断テストでは合計 26 分，授業の約 1/3 の時間を費やしてしまう．したがって，制限時間を減らし，少ない解答数で学習者の素養を診断しなければならない．本研究では従来の素養診断テストの課題として下記の 4 点が課題であると検討した．

- 問題の難易度が全体的に低く，難易度に差がない
- 難易度に関わらず，1 問 1 点
- 問題数が少ない
- 制限時間が長い

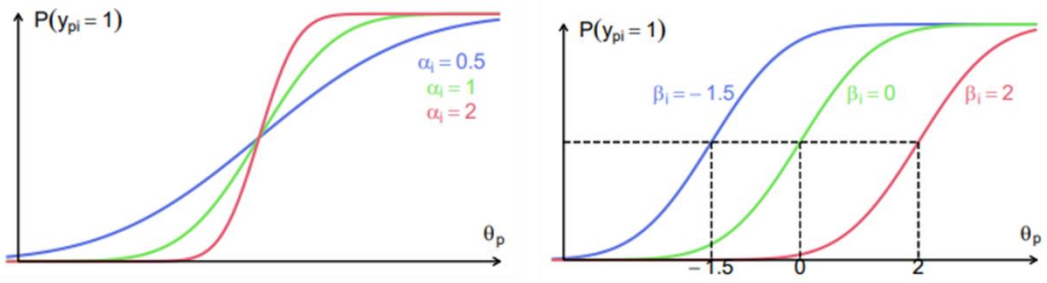
3章　プログラミング素養を測る手法の提案

3.1　項目応答理論（IRT）

異なる複数のテストによる測定結果を相互に比較可能にするためには，異なるテストの結果が基準上の値で表現される必要がある．項目反応理論は，このような共通尺度を構築する際に非常に有用なテスト理論であり，実用的な観点からも，高品質なテストの作成，実施，および運営に大きく貢献する．資格試験・能力試験などの様々なテストの開発・評価に活用されているテスト理論．項目応答理論の利点は以下のようである[5]．

1. 項目の困難度が受験者の集団とは独立に定義される
2. 受験者の特性尺度値が回答した項目群とは独立に定義される
3. 項目の困難度と受験者の特性尺度値とが同一尺度上に位置づけて表現される
4. 測定精度が特性尺度値の関数として表され，受験者個人ごとにきめ細かい測定精度の評価が可能になる

よって，受験者を評価したテストに対する受験者や受験者の解答データを確認することで，そのテストの難易度や問題の適正度を調べることができる．これによって素養診断テストのテスト項目が受験者の能力を適正に評価できるかを示す指標をつくることができる．図 3-1 は IRT 曲線の例で，横軸 θ が被験者の能力の値で，縦軸がその θ の能力を持った被験者が正答する確率を表した図である[6]．図 3-1(a)は項目識別力 α の値が大きいほど正誤の識別が大きいことが分かる．図 3-1(b)は項目困難度 β の値が大きいほど問題が難しいことがわかる．従来の素養診断テストは，図 2-8 から分かるように問題の難易度が低いテストであることが分かる．素養診断テストの適性度を調べ，本研究の手法から指標を作成する．



(a)項目識別力 α　　(b)項目困難度 β

図 3-1　IRT 曲線

3.1.1　項目応答理論と古典的テスト理論

測定結果は従来，古典的テスト理論と呼ばれる正答数得点や偏差値などが一般的である．しかし，同一受験者でも問題項目が異なると得点が変動し，また，特定の能力が同じでもテストの難易度が異なる場合には得点も異なるという問題がある．実際のテストでは，同一年度に複数回実施され，異なるテストによる測定結果を相互に比較する必要がある．異なるテストによる測定結果を相互に比較可能にするためには，それらの結果が共通の尺度上で表現される必要がある．

項目反応理論は，このような共通尺度を構築する上で非常に有益なテスト理論であり，実用的な観点からも高品質なテストの開発，実施，運営に役立つ．テストは複数の項目から構成される場合が多い．項目応答理論ではテストの最小単位を項目と呼び，m個からなるテストがあるとしたとき，j番目の項目への被験者の正答は 1，誤答は 0 で表すと

$$u_j = \begin{cases} 1, & \text{項目}j\text{に正答した場合} \\ 0, & \text{項目}j\text{に誤答した場合} \end{cases}$$

という 2 つの値となる．これを項目得点と呼ぶ．このときテスト得点を U とすると

$$U = \sum_{j=1}^{j=J} u_j$$

となる．一方，項目応答理論では，u_jを取り扱い，テスト得点に相当する量を計算する．

3.1.2　項目応答モデル

項目応答理論では項目特性関数のモデルとしていくつかの未知のパラメタをもつ数学的な関数を仮定し，項目反応データに基づいてそのパラメタを推定する方法をとり，次の 2 つのパラメタを用いる[7]．

1. 被験者母数：被験者の特性値
2. 項目母数：j番目のテスト項目の特性値
   - 項目識別力a_j：被験者特性値の違いが正答確率のどの程度敏感に反映するかを示す．
   - 項目困難度b_j：項目の難しさを決める
   - 当て推量c_j：実力ではなく全く正解できない被験者が偶然正答してしまう確率を表す．

被験者の特性値を θ とすると，横軸に潜在特性 θ，縦軸に正答確率を配する．横軸が潜在特性であるため，具体的な関数で表せれない．そこで標準正規分布の密度関数

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

の累積分布関数

$$\Phi(f(\theta)) = \int_{-\infty}^{f(\theta)} \varphi(z)dz\,,\ f(\theta) = a_j(\theta - b_j)$$

を用いる．これを正規累積モデルという[8]．正規累積モデルは積分を含んでおり，取り扱いを複雑になってしまう．そこでロジスティック分布関数を利用した近似公式を用いる．

$$\int_{-\infty}^{f(\theta)} \varphi(z)dz \cong \frac{1}{1 + \exp\left(-D \times f(\theta)\right)}$$

D は尺度因子であり，D=1,7 のときに θ の全域にわたって食い違いが 0.01 以下になることが分かっている．

3.1.3　被験者母数の推定

項目応答理論においては，多くのパラメタが存在し，これらの推定はロジスティックモデルにおいても必ずしも容易ではない．被験者の母数 θ を推定するためには，既にさまざまな手法が提案されているが，その中でも代表的な最尤推定法とベイズ推定法を説明する

i．最尤推定法

尤度関数と呼ばれる確率分布関数を最大化するようなパラメタを求める手法．また，多くの場合，尤度関数よりもその対数をとった対数尤度関数の方が簡単な形をしているため対数尤度関数の最大値エお与えるパラメタを求めることになる．項目モデルの場合，尤度関数は

$$L(u_i|\theta_i) = \prod_{j=1}^{n} p_j(\theta_i)^{u_{ij}}\, q_j(\theta_i)^{1-u_{ij}}$$

であり，対数尤度関数は

$$L(u_i|\theta_i) = \sum_{j=1}^{n} [u_{ij} \log p_j(\theta_i) + \left(i - u_{ij}\right) \log q_j(\theta_i)]$$

と表せる[7]．

対数尤度関数をθ_iで

$$L'L(\theta_i) = \frac{\partial}{\partial \theta_i} \log L(\frac{u_i}{\theta_i})$$

のように偏微分する．偏導関数の値を 0 とおいて，θ_iに関して方程式を解いて最尤推定値θ_iを求める．

ii．ベイズ推定法

ベイズ推定法は尺度値の分布を用いて尺度値を推定する方法である．最尤推定法では全問正解や全問誤答の被験者の尺度値の推定値は推定できない．ベイズ推定を用いると，それらの被験者の尺度値が推定できる．被験者母数をθ_iとし，その条件付き分布 $f(\theta_i|u_i)$ はベイズの定理より，表される[8]．

$$f(\theta_i|u_i) = \frac{f(\theta_i)f(u_i|\theta_i)}{f(u_i)}$$

ベイズ法に用いられる推定は事後分布の最頻値を推定値とする MAP 推定や事後分布の平均値を推定値とする EAP 推定がある．本研究では MAP 推定法を用いて被験者母数を推定する．

3.1.4　項目母数の推定

2 項目母数の項目反応理論において，通常は被験者母数の推定も同時に行う必要があるため同時最尤推定法が提案されたが，数理統計学的に好ましくない性質を有している．特に，データを増やすと推定値が不安定になり，信頼性が低下するという問題がある[7]．そのため，近年では同時最尤推定法はあまり採用されておらず，現在の主流は周辺最尤推定法である．この手法では，反応パタンが与えられた時の項目母数だけの尤度関数を構成し，推定を行う方法である．この方法では原理的に項目母数の推定のみ行えばよい[10]．

項目母数の推定は多くの場合 EM アルゴリズムが利用され，項目母数の計算は項目反応理論において最も計算量が多い．本研究では 2 母数ロジスティックモデルに対応した Easy Estimation[11]と呼ばれるソフトウェアを利用し，被験者母数と項目母数の推定を行った．

3.2　素養診断のコンピュータ適応型テスト（CAT）

3.2.1　　適応型テスト

項目の困難度をまんべんなく一様に含むテストは全ての被験者に対してほぼ等しい精度の測定を実施することができるが，十分に満足できる精度が得られるとは限らない．また，困難度に差がないテストは一部の被験者に対しては十分に満足できる精度の測定が可能であるが，他の被験者に対して不十分な測定となってしまう．そこで，被験者個人ごとに最適な困難度の特性を持つ項目を選択しテスト出題することで，全ての被験者に対して十分に満足できる精度の測定が可能になる．コンピュータを使用し，項目応答理論によって事前に特性値が算出されているテスト項目を各受験者の応答を適時判断しながら出題し，効率よく受験者の能力推定値を算出するテストをコンピュータ適応型テストと呼ばれている[12]．使用されている例として SPI（総合適性検査）や CASEC（Computerized Assessment System for English Communication）などがある．

適応型テストでは被験者の測定に最適な項目を選択するのに図 3-2 のように直前に実施した項目に対して正答した場合，次の項目の難易度が高くなり，誤答した場合には，次の項目にはより難易度の低くなるという処理を逐次繰り返す．

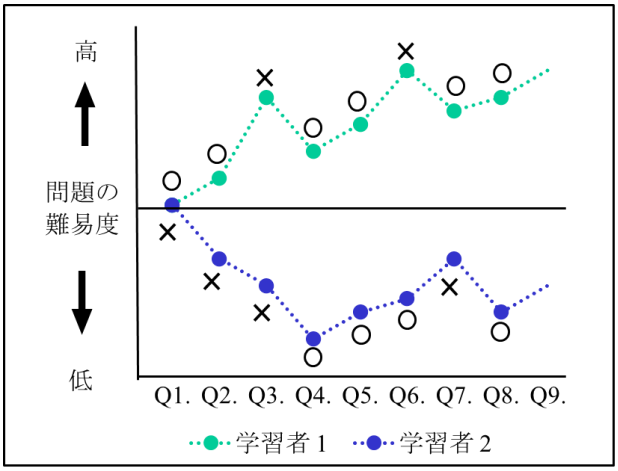


図 3-2 適応型テストの説明図

また，適応型テストの利点としては以下などが挙げられる[9].

- 全ての被験者に対して高い精度の測定の実施
- 精度を下げることなく被験者 1 人当たりに実施する項目数の低減と所要時間の短縮
- 難しい項目が続いて被験者にフラストレーションや不安を起こさせたり，易しすぎる項目が続いて飽きさせたりすることがない

実際の測定場面で適応型テストを実施するための具体的な方法については，図 3-3 に示すように分類される．本研究では，項目固定型多段階テストの多層構造を用いて素養診断テストを作成した．多層構造のテストは，被験者の項目に対する反応に基づいて困難度の層を上下して問題が出題されるテストである．

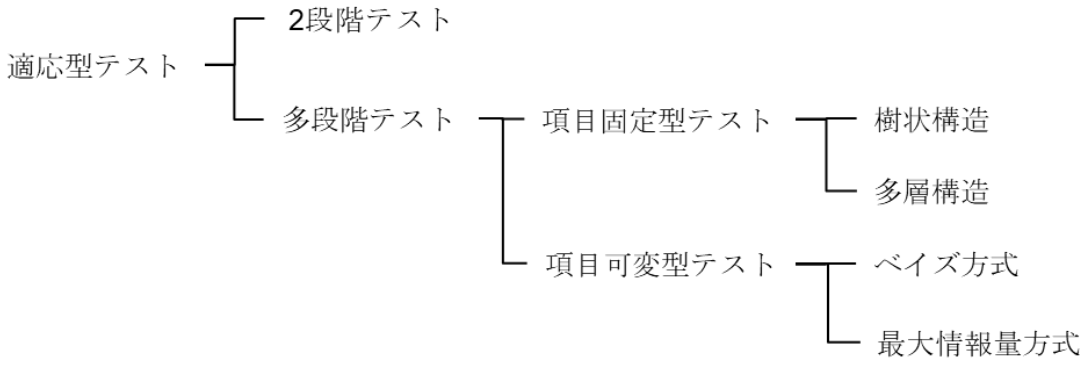


図 3-3　適応型テストの種類[9]

3.2.2　CAT 素養診断テスト

本研究では CAT を導入して，受験者の応答を適宜判断しながら出題する問題の難易度レベルを変えていくことで，従来の素養診断テストの課題であった 4 点について改良できると検討し，CAT 素養診断テストを作成した．従来の素養診断テストには「代入とシーケンス実行」「分岐繰り返し実行」「間違い探し」の 3 種類の問題があるが，本研究では代入とシーケンス実行問題のみ CAT を導入した．CAT の有用性を確認後に他の問題での導入を検討している．また，代入とシーケンス実行問題の命令の定義を図 2-1 から図 3-4 のように変更した．変更内容は加算・減算の処理とプログラミングの for 文のような繰り返しの処理を追加したことである．これによって，問題難易度に差が生まれる．

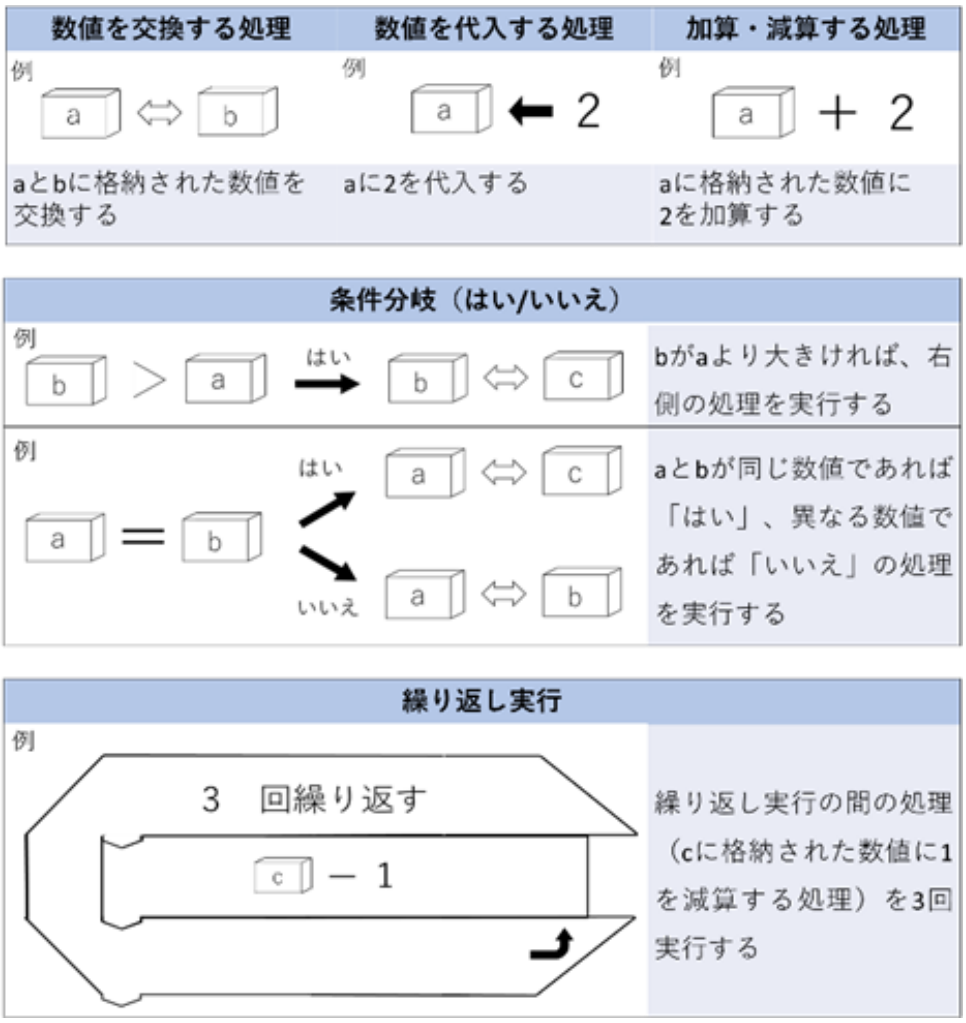


図 3-4　代入とシーケンス実行問題の処理の説明図

また，IRT によって推定する指標作成のデータが不足しているため，本研究の CAT 素養診断テストでは従来の素養診断テストの結果やプログラミング授業の講師の意見，Web アプリの動作確認に協力した学生の結果などを参考にそれぞれの処理に重みを付けレベルごとの点数を設定した．難易度はレベル 1～9 まで作成し，レベルごとの問題の詳細は表 3-1 に示す．問題例として，図 3-5 ~ 3-7 にレベル 1，5，8 を示す．

表 3-1 難易度ごとの問題内容と点数の重み

難易度	段数	問題内容	点数
1	2	数値の交換と代入，加算・減算	2
2	3	数値の交換と代入，加算・減算	3
3	3	数値の交換と代入，加算・減算 条件分岐	5
4	3	数値の交換と代入，加算・減算 繰り返し実行	7
5	4	数値の交換と代入，加算・減算 条件分岐	8
6	4	数値の交換と代入，加算・減算 繰り返し実行	10
7	4	数値の交換と代入，加算・減算 条件分岐・繰り返し実行	12
8	5	数値の交換と代入，加算・減算 条件分岐・繰り返し実行	15
9	5	数値の交換と代入，加算・減算 条件分岐・繰り返し実行（条件付き）	18

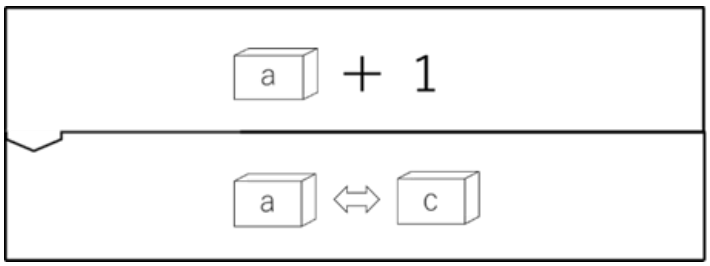


図 3-5 レベル 1 の問題例

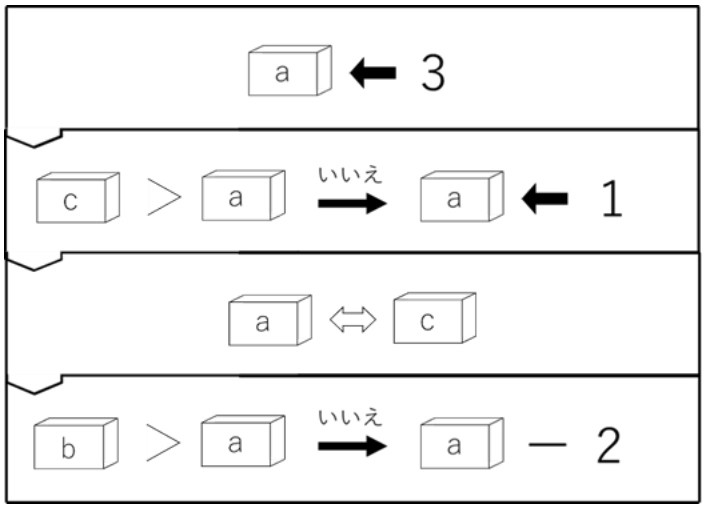


図 3-6 レベル 5 の問題例

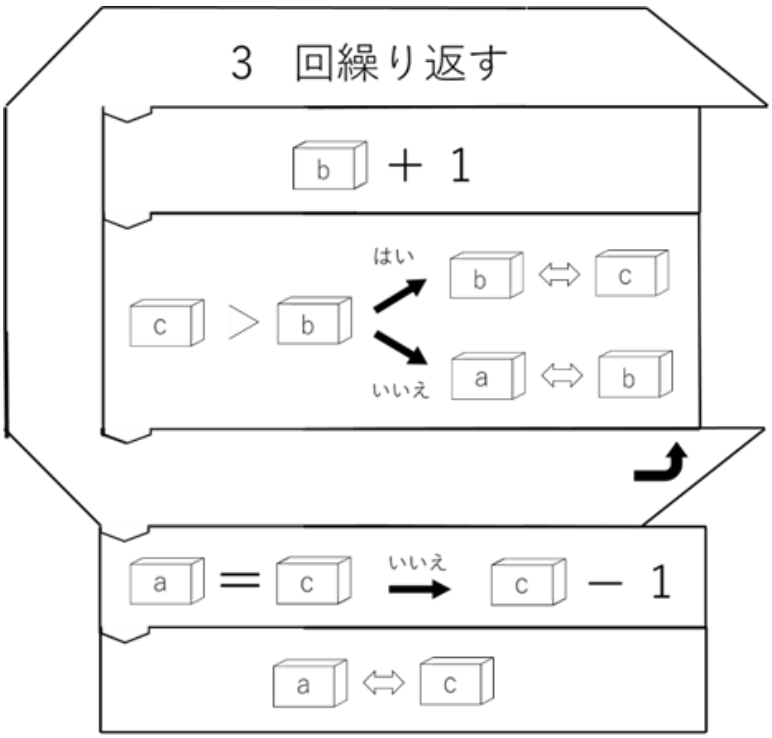


図 3-7 レベル 8 の問題例

4章 検証実験と結果

4.1 実験概要

4.1.1 目的

従来の素養診断テストの一問一答形式を変更し，本研究で提案している CAT のプログラミング素養診断テストを用いる．本来の用途として想定しているプログラミング初学者を対象に CAT のプログラミング素養診断テストの検証実験を行う．従来のプログラミング素養診断テストと比較し，従来のプログラミング素養診断テストでは選別できていなかった学習者を識別できているか調べ，CAT のプログラミング素養診断テストの有用性を検証する．

4.1.2 検証実験の実施内容

三重大学工学部総合工学科電気電子工学コースの科目であるデータサイエンスⅡの 2023 年度の受講者であり，かつ今回の研究対象である新規にプログラミングを学習する 87 名を対象にして検証実験を行った．過年度生は全員が再履修者でプログラミングが既習であるため対象外とした．従来の素養診断テスト（一問一答形式の代入とシーケンス実行問題 6 問と間違い探し問題 9 問）と本研究の CAT 素養診断テストを実施した．CAT 素養診断テストは図 4-1 のように Web アプリを作成し，問題ごとに制限時間を設けて実施した．CAT 素養診断テストの概要を下記に示す．

問題種類 ： 代入とシーケンス実行問題
難易度 ： レベル 1 ～ 9
点数 ： 各レベルの配点の合計
問題数 ： 全 90 問，各レベル 10 問
制限時間 ： 10 分 （約 20 問解答想定）

本研究で得られたデータと演習の進捗や科目の成績を比較し，CAT によって素養診断テストの高得点者がプログラミングに関係する科目の成績が高いまたはプログラミングの能力が高い学習者に診断できているか，CAT 素養診断テストの正当性を確認する．

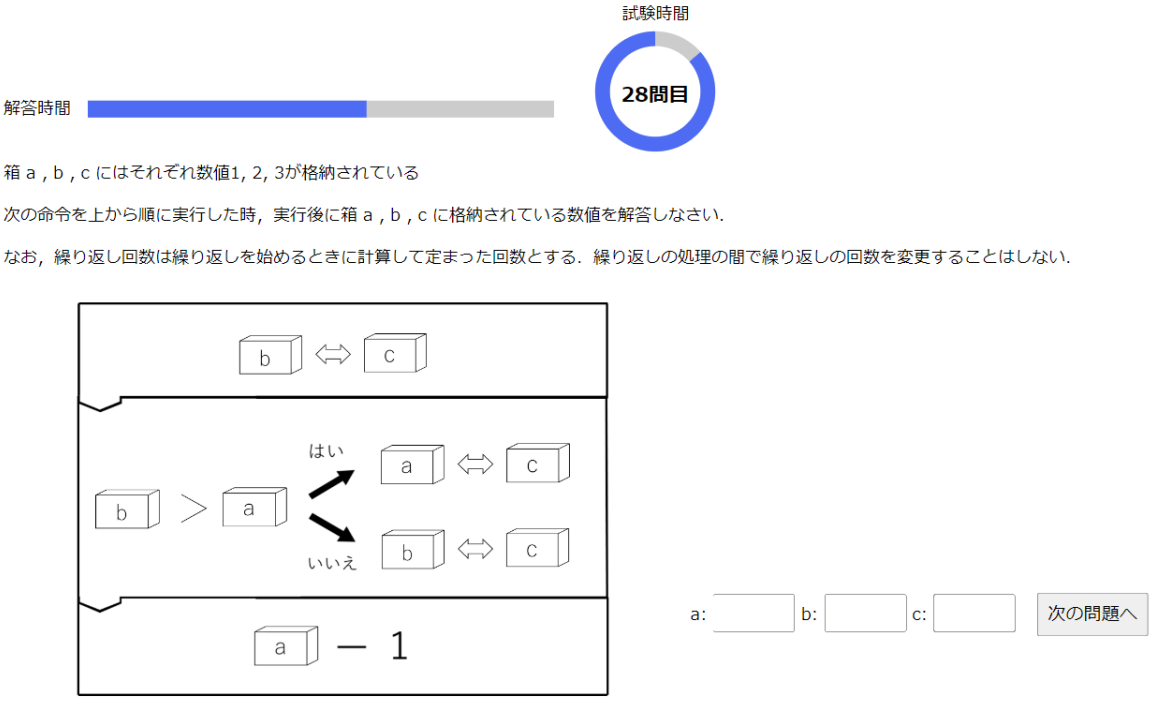
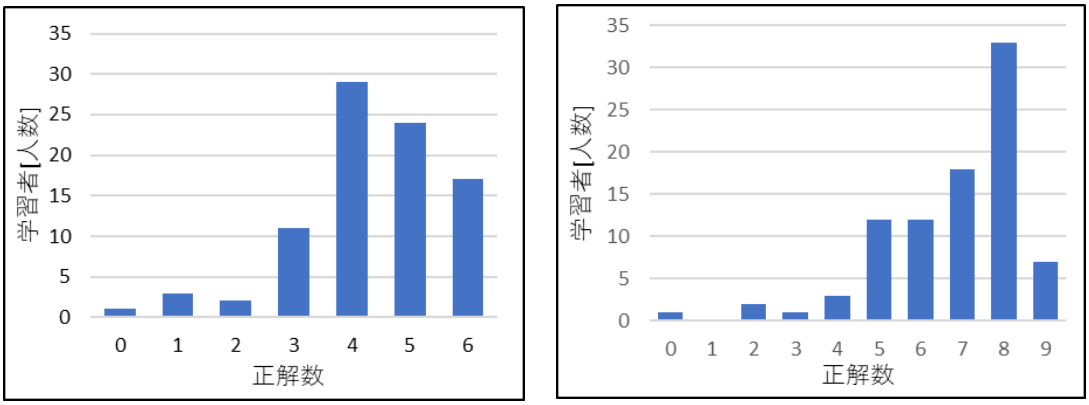


図 4-1 CAT 素養診断テストの受験中画面

4.2　従来の素養診断テストの結果と考察

4.2.1　　基本的統計量

従来の素養診断テストでは代入とシーケンス実行問題 6 問と間違い探し問題 9 問を実施した．それぞれのテストの得点分布を図 5-1 に示す．代入とシーケンス実行問題は平均正答率が 0.72 で間違い探し問題は 0.75 であった．従来の素養診断テストの課題でも述べたように，高得点者に分布の偏りが生じていること分かる．



(a)代入とシーケンス実行問題　　(b) 間違い探し問題

図 5-1 従来の素養診断の結果（今年度）

4.2.2　項目母数の推定

2 母数ロジスティック推定を用いて，代入とシーケンス実行問題と間違い探しのそれぞれのテストの項目母数a_j（項目識別力）とb_j（項目困難度）を推定した．それぞれ表 5-1 に示す．この表から，テストの各項目の特徴やテストの適性度を把握することができる．a_jは値が大きいほど能力の有無によって正答率に差がつき，識別力ある項目とわかる．b_jは 0 を基準に値が大きいほど困難度が高い問題で，小さいならば困難度が低い問題となる．

代入とシーケンス実行問題は問題が進むにつれて問題の難易度が上がっていくようにテスト作成されており，代入とシーケンス実行問題のb_jの値から問題が進むにつれて困難度が大きくなっていることが分かる．

間違い探し問題の難易度はほぼ一定として作成されているが推定された困難度は問題 4 と 6 だけ正の値を示している．また，問題 4 は極端に困難度の高い問題だと推定されており，識別力も低く不適切な問題だと分かる．問題内容を確認すると，ひっかけ問題のような内容になっているため，問題として不適切だったと考えられる．このように項目母数を推定し，問題を調節することで，より信頼性の高いテスト実施ができる．

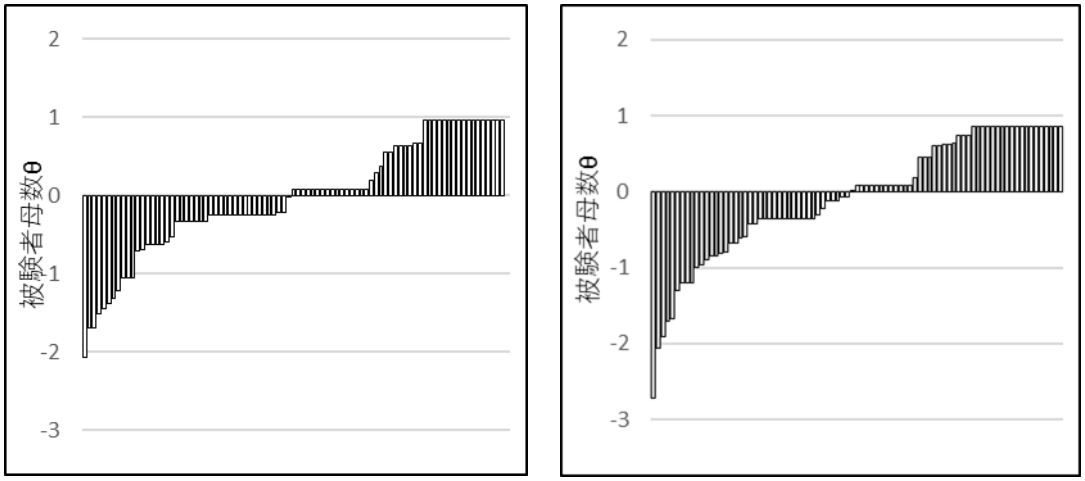
表 5-1 代入とシーケンス実行問題と間違い探し問題の項目母数の推定

問題番号	代入とシーケンス実行問題		間違い探し問題	
	a_j	b_j	a_j	b_j
1	0.378	-2.503	1.671	-1.199
2	0.657	-2.771	3.599	-1.415
3	1.663	-1.413	3.081	-1.113
4	0.530	-1.396	0.150	8.679
5	0.420	-0.841	0.687	-3.261
6	1.202	0.562	1.307	1.659
7			0.697	-1.761
8			2.980	-1.228
9			0.999	-1.065

4.2.3　被験者母数の推定

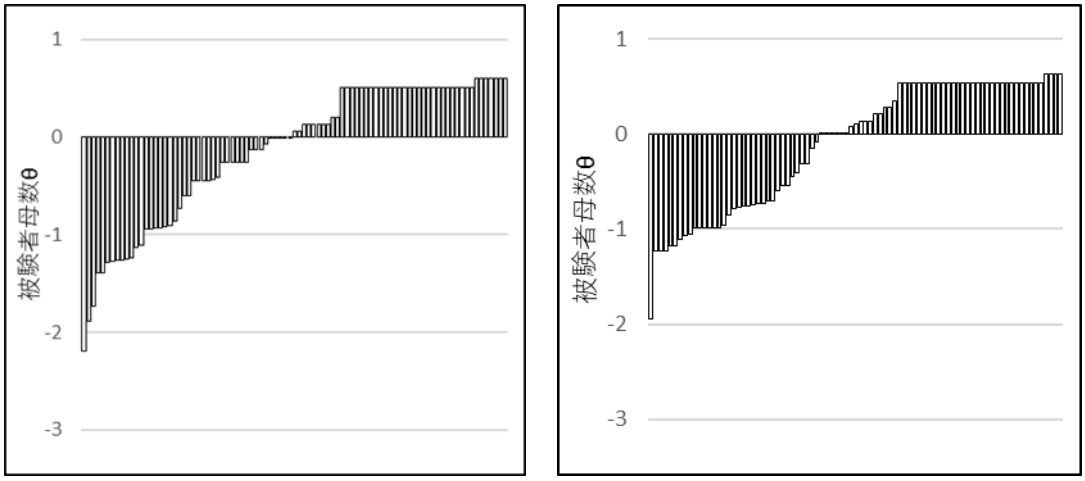
次に，MAP 推定法を用いて，今年度と 2022 年度実施（85 名）の代入とシーケンス実行問題と間違い探し問題の学習者の特性値である被験者母数θを推定した．被験者母数θは 0 を基準に正の方向に大きいほど被験者の能力が高く，負の方向に大きいほど被験者の能力が低い．図 5-2，5-3 にそれぞれのテストから得られた学習者の被験者母数θを昇順に並べた分布を示す．

図 5-2 の代入とシーケンス実行問題では今年度と 2022 年度実施を比較すると，同様な分布を示しており，θが小さい学習者の識別力は高いが，θが 0 以上の分布では識別力が低いことが分かる．また，図 5-3 の間違い探し問題でも今年度と 2022 年度実施を比較すると，同様な分布を示しており，θ が小さい学習者の識別力は高いが，θが 0.5 に近い学習者の識別ができていないことが分かる．



(a)今年度実施　　(b) 2022 年度実施

図 5-2　代入とシーケンス実行問題



(a)今年度実施　　(b) 2022 年度実施

図 5-3　間違い探し問題

4.3　CAT 素養診断テストの結果と考察

4.3.1　　基本的統計量

本研究で実施した CAT 素養診断テストの得点の分布を図 5-4 に示す．テストスコアの平均は 128.9，平均解答数は 21.9，平均正解数 17.5，平均正答率 0.799 であった．得点分布は正規分布に近い得点分布をしており，低得点者と高得点者を識別できていることがわかる．

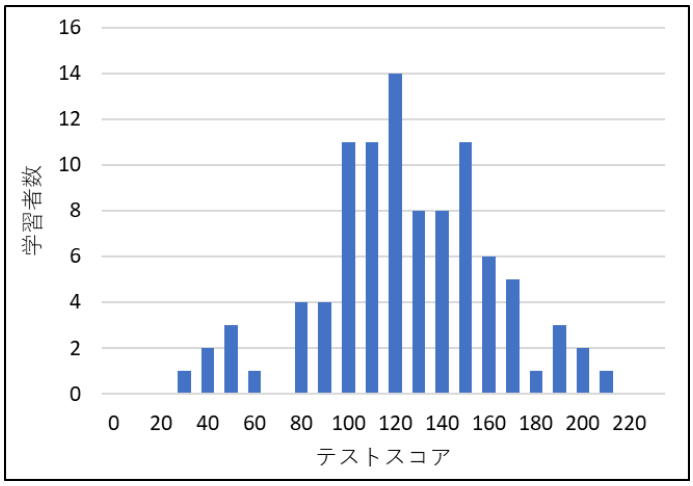


図 5-4 CAT 素養診断テストの得点分布

次に,CAT 素養診断テストで学習者が到達した最高レベルの分布を図 5-5 に示す．図 5-5 から高レベルに偏りが生じており，図 5-1(a)に近い分布を示しているが，学習者によって解く問題数や正解数が異なるため．到達レベルが同じでもテストスコアに差が発生することが分かる．したがって，CAT 素養診断テストの高得点者の識別度向上が考えられる．また，表 5-2 にレベルごとの解答者数や解答数などの内訳を示す．

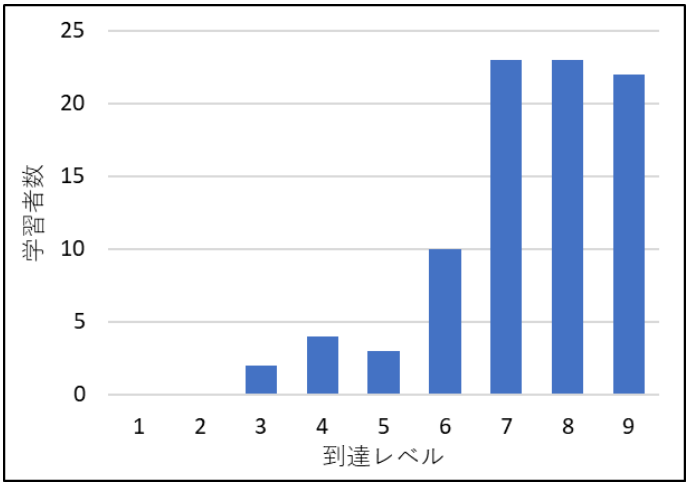


図 5-5　CAT 素養診断テストの到達レベル

表 5-2　CAT 素養診断テストのレベルごとの内訳

レベル	解答者数	全正答数	全解答数	正答率	重み
1	87	215	234	0.919	2
2	87	211	253	0.834	3
3	87	220	271	0.812	5
4	85	186	202	0.921	7
5	81	194	214	0.907	8
6	78	183	213	0.859	10
7	68	170	222	0.766	12
8	45	96	148	0.649	15
9	22	48	69	0.696	18

4.3.2　最終試験や成績などとの比較

従来型の一問一答形式と本研究の CAT 形式を最終試験や成績などと比較した結果を表 5-3 に示す．従来型の代入とシーケンス実行問題の満点者が 17 人であったため，CAT 素養診断テストの上位 17 名と比較した．CAT 素養診断テストの最高点は 214 点で上位 17 人目は 160 点であった．　従来の素養診断テストでは満点を取ると，素養に差があっても同程度の素養として認識されるが，CAT 素養診断テストによって，高得点者の識別が可能となった．代入とシーケンス実行問題の高得点者の中で科目の最終の評価 9，10 を取得したのは一問一答形式だと 41.1%，CAT 形式だと 53.0%となっており，精度が向上したことがわかる．CAT 素養診断テストで高得点であったが評価が低かった学習者を確認すると，欠席や小テスト未実施など真面目に取り組んでいないため素養はあるが不合格者になった学習者が 87 名中 1 名いた．また，CAT 素養診断テストが低得点で評価が高かった学習者が 87 名中 4 名いたが授業期間でプログラミング能力が向上したと考えられるため誤差の範囲内だと考えた．

表 5-3 従来形式と CAT 形式の比較結果

	代入とシーケンス実行問題			
	一問一答	CAT 形式		
	満点（6 点）	160 点以上	170 点以上	180 点以上
事前素養診断（高得点者）	17	17	12	7
事後判定（評価 9，10）	7	9	7	3

5章 プログラミング素養診断テストの課題

本研究の CAT 素養診断テストは項目固定型テストであり，ベイズ方式や最大情報量方式の項目可変型テストを実現するための項目データ量が足りていないことが課題である．また，項目可変型テストには項目バンクが不可欠である[14]．項目バンクには相当数の項目を貯蔵する必要があり，項目の自動生成が必要である．また，項目を作成するだけでなく，全ての項目にはその項目の項目母数（項目困難度，項目識別力）の情報を登録する必要がある．

解決方法としては，新たに素養診断テストを受験する学習者のテスト結果を項目の特性値を推定するためのデータとして蓄積する．これによって，特性値の推定を更に精度の高い特性値を得ることが可能である．また，項目や被験者が少ない場面ではノンパラメトリック IRT（NIRT）モデルが有用である．NIRT モデルの代表的なものとしてモッケン尺度分析がある．一般的な IRT モデルで行われるような個々の項目，被験者に関する母数が推定されることは無く，項目間の関係から尺度がその持つべき望ましい性質を備えているかを検討している[15]．そのため，素養診断テストに NIRT を用いることで課題が解決できる可能性があると考える．

6章 まとめ

プログラミングの授業において，プログラミングの素養によってプログラム作成にかかる時間には大きな差がある．教師はプログラミングに苦労する学習者を中心に指導するため，プログラミングを得意な学習者が時間を余らしてしまうのが現状である．この問題を改善するために本研究室ではプログラミング素養診断テストの研究が進められてきた．プログラミング素養診断テストは授業を受ける前に学習者の学習可能性を把握するために素養の有無を事前に判定するテストである．

従来の素養診断テストではプログラミング素養が低い学習者の識別は可能であったが，高得点者での識別が困難であった．そのため，本研究ではコンピュータ適応型(CAT)素養診断テストを Web アプリとして開発した．プログラミング素養診断テストの識別度向上のために項目応答理論とコンピュータ適応型テストを用いて，プログラミング初学者を対象に CAT 素養診断テストを実施した結果，従来では識別できていなかった素養診断テストの高得点者の識別が可能になった．

CAT 素養診断テストの今後の課題点として，IRT に必要なデータ量が足りない事や項目数がすくないことが考えられる． 解決案としては問題の自動生成ツールの開発や受験者の解答後に特性値を推定するデータとして蓄積するシステム構築，NIRT モデルを用いた素養診断テストの作成が考えられる．

謝辞

本研究の進行及び作成に当たり，懇切丁寧なご指導と御督励を賜った本学工学研究科電気電子工学専攻の北英彦准教授，高瀬治彦教授，川中普晴教授に感謝いたします．また，日頃熱心に討論していただいた計算機工学研究室，情報処理研究室の皆様方に厚く御礼申し上げます．最後に本論文をまとめるにあたり，助言，討論，その他お世話になったすべての方々に感謝いたします．

参考文献

[1] 高桑稔，北英彦：プログラミング能力向上を目的としたプログラムテストの学習システム,CIEC コンピュータ利用教育学会,PC カンファレンス 2014（2014）

[2] 小林史生，北英彦：学習者のプログラミングの素養を調査する手法，CIEC コンピュータ利用教育学会，PC カンファレンス 2014（2014）

[3] 寺久保丞，プログラミング素養診断テストの有用性の調査に関する研究，2015

[4] 大津悠，プログラミング素養診断テストの弁別度向上に関する研究，2016

[5] 野口裕之，斉田智里，孫媛．項目応答理論の基礎と応用，The Annual Report of Educational Psychology in Japan 2005，Vol.44，32-36，2005

[6] 分寺杏介．「統計的方法論特殊研究（多変量解析）」．2023

[7] 豊田秀樹．項目反応理論[理論編]．朝倉書店，2005.

[8] 豊田秀樹．項目反応理論[入門編]【第 2 版】．朝倉書店，2012.

[9] 芝祐順．項目反応理論　基礎と応用．東京大学出版会，1991.

[10] 古谷博史，愛甲弥生．項目反応理論を用いたプログラミングテストの分析，宮崎大学工学部紀要，巻 36，p．333-338，2007

[11] 熊谷龍一.2009　初学者向けの項目反応理論分析プログラム EasyEstimation シリーズの開発，日本テスト学会誌，5，107-118.

[12] 中村 洋一．コンピュータ適応型テストの可能性．日本語教育，148 巻，p．72-83．2011

[13] 彦坂 知行．多人数でのプログラミング演習における 学習者のコーディング状況の 把握システムに関する研究，CIEC 研究会報告書，7 巻，p18-24，2016

[14] 伊藤祐郎.項目バンクによって広がるテスト開発の可能性,日本語教育 148 号，p．57-71，2014

[15] 豊田秀樹．項目反応理論[中級編]．朝倉書店，2013.

発表実績

[1] Ryo Araki and Hidehiko Kita : Development of the System to Identify Programming Learning Potential, The 13th International Symposium for Sustainability by engineering at Mie University (Research Area C), A-5, 2023

[2] 荒木諒, 北英彦:プログラミングの学習可能性を把握するための素養診断テスト, 2023 九州 PC カンファレンス論文集, pp.9-12, 2023.