

厚生労働科学研究費補助金
(政策科学総合研究事業（臨床研究等 ICT 基盤構築・人工知能実装研究事業）)

総括研究報告書

ICT・AI 技術等を活用した看護師等国家試験問題作問システムの検討と
その実装に向けた研究

研究代表者 聖路加国際大学大学院看護学研究科 林 直子 教授

研究要旨

当研究班が行った先行研究（令和 4～6 年度：保健師助産師看護師の問題作成の支援と効率化に向けた ICT・AI 技術等の活用策の検討のための研究、研究代表者林直子）では、ICT・AI 技術を活用した看護師等国家試験問題作問システム（以下作問システム）を検討し、さまざまな成果を得た。そこで本研究では、先行研究で開発した必修問題の作問システムのプロセスを適用した一般問題の作問システムを開発すること、及び試験委員によって作成された問題の校閲と出題形式（Taxonomy）を評価するシステム（校閲システム）を検討すること（研究 1）、さらに厚生労働省医政局看護課との協働により実際の試験委員会委員を対象にルールブックの利用可能性と作問システムの実装可能性を評価すること、国内外の他資格の国家試験における ICT・AI 活用状況と機密性担保に関して情報を整理し、今後の詳細な調査や議論の方向性を検討すること（研究 2）を目的とした。初年度の令和 7 年度は、研究 1、研究 2 それぞれ以下の成果を得た。

研究 1 では、看護師国家試験の一般問題出題数が多い「人体の構造と機能」「疾病の成り立ちと回復の促進」「基礎看護学」「成人看護学」の 4 科目を対象に、AI を活用した作問システム開発を進めた。厚生労働省から入手した過去 10 年分の国家試験問題 640 問と模擬試験問題 600 問を基礎データとして用いた。このうち、国試過去問については各科目の出題年ごとの平均正解率を基に閾値を設定し、正解率と識別指数を用いて問題の良否を 4 種類に分類、可視化した。また、「人体の構造と機能」と「疾病の成り立ちと回復の促進」の 2 科目について、類似問題分析を行った。その結果、必修問題と比べて類似問題数がきわめて少ないことが明らかとなり、必修問題のようなルール抽出は困難である可能性が示唆された。今後「基礎看護学」「成人看護学」についても同様の分析を行い、作問ルールブック作成とシステムへの活用を検討する。

さらに、誤答選択肢生成（作問システム開発）においては、生成 AI の temperature 設定を調整し複数候補を集約することで、多様性の高い誤答選択肢生成を試みた。加えて、LLM-as-a-Judge による自動判定を導入し、不適切な選択肢を除外した結果、実際の国家試験問題と同程度の多様性と品質を備えた誤答選択肢を自動生成できることが確認された。次年度は、実際の作問現場での有用性を検証する予定である。

研究 2 では、先行研究で作成した必修問題作問ルールブックを、実際の国家試験委員会に於いて試験委員が活用することに対する受容性（作問の負担感への影響、有用性等）を検討した。そ

の結果、Excel 形式で提供したルールブックは、作問のヒントや作問会議の検討支援に有効であり、作問者の自由な発想を妨げることなく、作問作業の負担軽減にもつながるとの評価が得られた。一方、情報量の充実や継続的な更新の必要性も示された。次年度も継続的に調査を行いより広く意見を収集することで、ルールブックの利用可能性についてさらに検討する予定である。

また国内外の他資格・分野の国家試験等の ICT・AI 活用状況並びに機密性の担保について、宮本らが行った先行研究（2023）より、国内他資格、外国の看護師資格試験における CBT の活用状況に関する、制度・仕組みの全体像は、既に十分把握できていることが確認された。今後の本課題における調査、検討の方向性として、(1)諸外国における看護師資格試験について、先行する調査以降の変更を調査すること、(2) (1)を踏まえ、すでに海外の CBT で実施されている方式を日本に適用する場合の検討事項、実現可能性を具体的にシミュレーションすること、の2点が挙げられ、次年度の課題とした。

研究分担者氏名 所属 職位

徳永 健伸 東京科学大学 教授
荒瀬 由紀 東京科学大学 教授
宇佐美 慧 東京大学 准教授
佐伯 由香 人間環境大学 教授
佐々木 幾美 日本赤十字看護大学 教授
米倉 佑貴 聖路加国際大学 准教授
佐居 由美 聖路加国際大学 教授

研究協力者氏名 所属 職位

山田 寛章 東京科学大学 助教
宮本 千津子 東京医療保健大学 教授
鈴木 久美 大阪医科薬科大学 教授
野崎 真奈美 順天堂大学 教授
水戸 優子 神奈川県立保健福祉大学 教授
伊藤 圭 大学入試センター 准教授
西崎 祐史 順天堂大学 教授
木村 理加 聖路加国際大学 助教
山本 真実 聖路加国際大学 助教

A. 研究目的

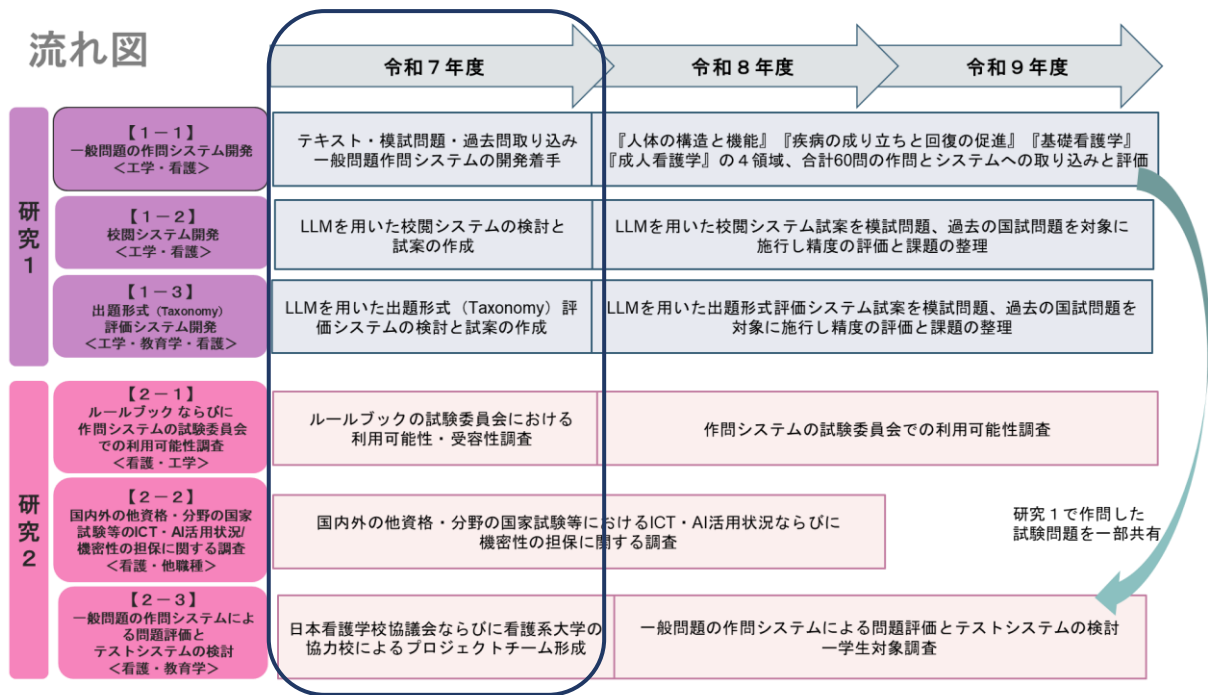
当研究班は、令和4～6年度の「保健師助産師看護師の問題作成の支援と効率化に向けたICT・AI技術等の活用策の検討のための研究」（研究代表者林直子）において、ICT・AI技術を活用した看護師等国家試験問題作問システム（以下作問システム）を検討し、さまざまな成果を得た^{1)~4)}。具体的には、①過去10年間の必修問題500問について正解率、識別指数に基づき問題を分類し、出題テーマに則した作問ルール（以下ルールブック）を作成、②6種類（7つの生成方法）の大規模言語モデル（Large Language Models: 以下LLM）を用いて必修問題の誤答選択肢自動生成システムを構築、作問委員の負担軽減や効率改善への寄与を提示、③LLMの支援を得て作成した誤答肢は人（試験委員等）が作成した誤答肢と正解率、識別指数の点で遜色がない可能性の示唆を得たことである。一方、今後の課題として、一般問題など必修問題以外の問題の作問におけるLLMの利用可能性、作成したルールブックの試験委員会委員の受

容性、LLMの活用による、問題作成プロセスの機密性の担保などいくつかの課題が明らかとなった。

そこで本研究では、先行研究で開発した必修問題の作問システムのプロセスを適用した、一般問題の作問システムを開発すること、及び試験委員によって作成された問題の校閲と出題形式（Taxonomy）を評価するシステム（校閲システム）を検討すること（研究1）、さらに厚生労働省医政局看護課との協働により実際の試験委員会委員を対象にルールブックの利用可能性と作問システムの実装可能性を評価すること、国内外の他資格の国家試験におけるICT・AI活用状況と機密性担保に関して情報を整理し、今後の詳細な調査や議論の方向性を検討すること（研究2）を目的とする。

本研究は、令和7年度から令和9年度の3年計画で実施する。令和7年度は図1に示す令和7年度の計画を遂行した。

図 1 本研究の3年間の全体計画



1. 【研究 1】 ICT・AI 技術を活用した看護師等国家試験問題作問システム開発

担当：(研究代表者) 林、(研究分担者) 徳永、荒瀬、佐伯、佐々木、佐居、(研究協力者) 山田、宮本、鈴木、野崎、水戸、木村、山本

B. 研究方法

本研究班は、看護師等国家試験委員経験者をはじめ、看護教育学、看護情報学、医学（医師国家試験 CBT 研究有識者）、工学（情報工学、システム設計）、教育学（教育測定学）の専門家が研究分担者・協力者として参与し、研究方法の妥当性と課題を議論し、適宜方法の修正等を図りながら進めた。

1. 研究 1-1 看護師等国家試験 一般問題の作問システム開発

1) 看護師国家試験（一般問題）過去問分析

本研究では、看護師等国家試験の一般問題出題数の上位 4 科目、即ち人体の構造と機能、疾病の成り立ちと回復の促進、基礎看護学、成人看護学に焦点を当て、作問システムの開発に向けた作業を開始した。

当研究班が行った先行研究の結果、多肢選択肢試験問題の作問プロセスのうち、看護師国試問題作成において誤答肢作成が最も困難であること、また大規模言語モデル (LLM) が最も有効に活用されることが示されたため、本研究においても誤答肢生成のための作問システムを開発することとした。開発のための基礎データとして、厚生労働省と看護師国家試験模擬試験を実施している予備校から、それぞれ以下のデータを入手した。

①厚労省医政局看護課より看護師国家試験過去問題 10 年分（第 105 回～第 114 回）

一般問題 4 科目（人体の構造と機能 120 問、疾病の成り立ちと回復の促進 120 問、基礎看護学 210 問、成人看護学 190 問） 計

640 問

②東京アカデミー（医療系予備校）より前年度までの模試で使用した問題

一般問題 4 科目（人体の構造と機能、疾病の成り立ちと回復の促進、基礎看護学、成人看護学）各 150 問、計 600 問

次に、先行研究と同様に①で入手した 4 科目 640 問について、問題の評価（良問：評価○、良問だが難問：評価②、あまり適切ではない問題：評価③、非該当（いずれにも該当しない）：評価△）を決定するため、正解率、識別指数について、良問とそれを判別する閾値を決定した。一般問題は必修問題と異なり、目標正解率が明示されていないことから、過去 10 年間の正解率の平均値を算出し、これに基づき正解率の閾値を決定することとした。識別指数については、先行研究と同様に 0.2 を閾値とした。

この基準を元に、まず各科目の問題評価分析を行い、次に類似問題（質問テーマが同様のもの）を抽出し、評価の相違からルールの抽出を試みることにした。

2) 一般問題作問システムの開発

多肢選択肢問題は、ある問題に対し、複数の選択肢から一つまたは複数の正解を選択する形式の問題である。記述回答式の問題に比べ、採点に要するコストが低く、少ない労力で学習者の習熟度を効果的に測定できるテスト形式として、教育分野で広く用いられている。しかしながら、多肢選択肢問題においては「問題文」「正解選択肢」「誤答選択肢」の三つを作成する必要がある、問題作成におけるコストの面には課題がある。特に「誤答選択肢」は問題に対して複数用意する必要がある、作成コストが高い。適切な誤答選択肢を自動生成することは、多肢選択肢問題を作成する上で大いに役立つと考えられる。そこ

で、当研究班が行った先行研究と同様に、本研究においても大規模言語モデル (Large Language Model: LLM) を用いて、誤答選択肢を作成した。看護師国家試験における一般問題 4 科目について、誤答選択肢の多様性確保に焦点を当て、妥当かつ実際の試験問題における誤答選択肢と同程度の多様性を持つ選択肢候補を生成する手法を提案するべく、作業を進めた。提案手法では、生成結果の多様性を高めるために temperature(出力の多様性レベルを示すパラメータ)を高い値に設定し、複数の temperature で得られた候補を集約することとした。加えて、非妥当な誤答選択肢を除外し、適切な誤答選択肢を抽出する目的で LLM-as-a-Judge を用いた。

本研究における提案手法の生成フローを図 2 に示す。本手法では、生成機構、排除機構、抽出機構の 3 つの機構を用いて誤答選択肢の生成を行った。なお、これら全ての機構について、モデルは Llama-3.1-Swallow-8B-Instruct-v0.5 を使用した。

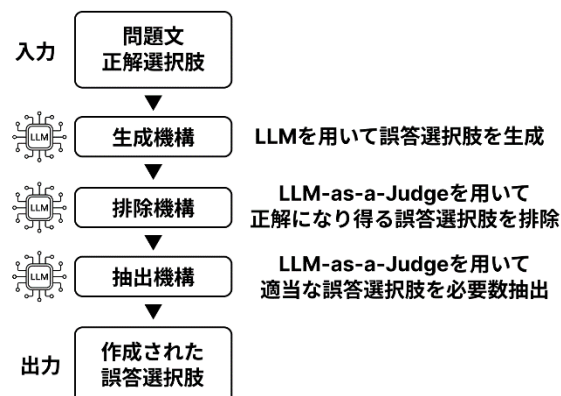
生成機構では、誤答選択肢を Five-Shot、ファインチューニング (SFT) の 2 つの手法でそれぞれ実際の誤答選択肢の数だけ生成させる。このとき、それぞれの手法について 3 種類の temperature を用いて生成させた。生成して得られたすべての誤答選択肢を集約し、これを次の排除機構に与えた。なお、temperature は (0.2, 0.4, 0.6) と (0.8, 1.0, 1.2) の 2 種類の組を試した。

排除機構では、生成機構より与えられた誤答選択肢それぞれについて、問題文に対する正解になってしまう可能性があるか LLM を用いて判定し、正解になりうると判定された誤答選択肢を排除した。

抽出機構では、与えられた誤答選択肢から、LLM が適切と判断したものを実際の誤答選択肢の数だけ抽出した。適切な誤答選択肢の基準は、システムプロンプトを通して与えた。基準は Rajiv によるガイドライン⁵⁾を日本

語に翻訳したものを基に作成した。

図 2 誤答選択肢の生成フロー⁶⁾



機構の三つの機構を用いて誤答選択肢の生成を行った。生成機構では、誤答選択肢を Five-Shot、ファインチューニング (SFT) の二つの手法でそれぞれ実際の誤答選択肢の数だけ生成させた。生成して得られたすべての誤答選択肢を集約し、これを次の排除機構に与えた。排除機構では、生成機構より与えられた誤答選択肢それぞれについて、問題文に対する正解になってしまう可能性があるか LLM を用いて判定し、正解になりうると判定された誤答選択肢を排除した。抽出機構では、与えられた誤答選択肢から、LLM が適切と判断したものを実際の誤答選択肢の数だけ抽出した。適切な誤答選択肢の基準はシステムプロンプトを通して与えるものとした。

2. 研究 1-2 校閲システム開発

看護師国家試験委員会により作成された問題の適切性、習熟度、職種・領域間での重複状況の確認を行うための校閲システムの検討と、試案の作成を行う計画である。対象とする 4 科目 (人体の構造と機能、疾病の成り立ちと回復の促進、基礎看護学、成人看護学) の出題問題について、(a) 類似した問題文を有する問題の検出、(b) 選択肢も含めた類似問題の検出モデルを構築する。テキスト表層のマッチングに加え、テキスト埋込に基づく

類似度の利用も検討する。類似性の判定根拠について工学系研究者では判断が難しいため、研究班内の看護研究者らと連携しながら取り組む。

3. 研究 1-3 出題形式 (Taxonomy) 評価システム開発

試験委員会が作成した問題の Taxonomy の妥当性を評価するシステムの試案を作成する。一般問題では Taxonomy I 型もしくは II 型を中心に出题することが求められるため、本研究では LLM による問題の解答過程に基づき Taxonomy を推定する手法を設計する。I 型に比較し II 型・III 型の問題は回答に高度な推論を要すると考えられる。そこで LLM が問題に解答するために辿る推論過程を明示的に出力させ、問題ごとの推論ステップの差異を考慮することで Taxonomy を推定する。この手法により看護師国家試験データに含まれる Taxonomy を正解とする分類モデルを構築する計画である。具体的には、問題文と選択肢を LLM に入力し、Taxonomy を予測する手法を開発する。

C. 研究結果

1. 研究 1-1 看護師等国家試験 一般問題の作問システム開発

1) 看護師国家試験 (一般問題) 過去問分析
厚労省から入手した 4 科目の過去問 10 年分について、科目ごとに正解率 (横軸)、識別指数 (縦軸) をプロットし、さらに各科目の「目標」別、「Taxonomy」別にプロットを作成した (添付資料)。正解率の閾値は、過去 10 年間の正解率の平均値 (科目順に 54.6%、53.4%、77.2%、70.8%) をもとに定めた (表 1)。この基準に基づき、各科目のプロットと評価を表 2~3 および図 3~6 に示した。

表 1 各科目の評価基準の正解率の閾値 (A)

科目	表 2 の A
人体の構造と機能	50%
疾病の成り立ちと回復の促進	50%
基礎看護学	75%
成人看護学	70%

表 2 評価基準

評価	正解率	識別指数
良問：○	A 以上	0.2 以上
良問だが難問：②	A 未満	0.2 以上
あまり適切ではない：③	A 未満	0.2 未満
非該当：△	A 以上	0.2 未満

表 3 問題評価結果

	人体の構造と機能 ※除外 1 問		疾病の成り立ちと 回復の促進		基礎看護学 ※除外 2 問		成人看護学	
評価	問題数	割合	問題数	割合	問題数	割合	問題数	割合
○	58	48.7%	49	40.8%	76	36.5%	78	41.1%
②	24	20.2%	40	33.3%	48	23.1%	51	26.8%
③	21	17.6%	19	15.8%	26	12.5%	31	16.3%
△	16	13.4%	12	10.0%	58	27.9%	30	15.8%
合計	119		120		208		190	

図3 人体の構造と機能
正解率－識別指数プロット

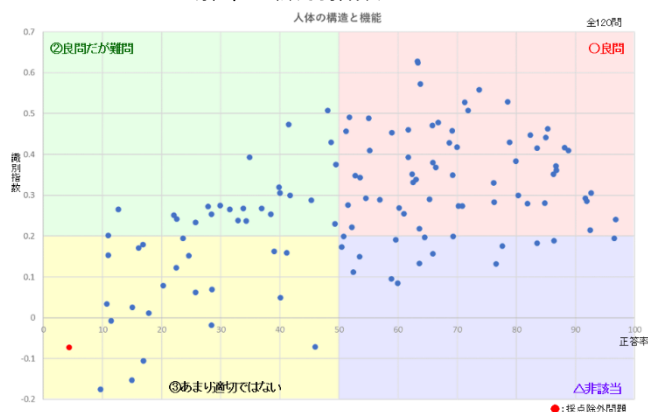


図4 疾病の成り立ちと回復の促進
正解率－識別指数プロット

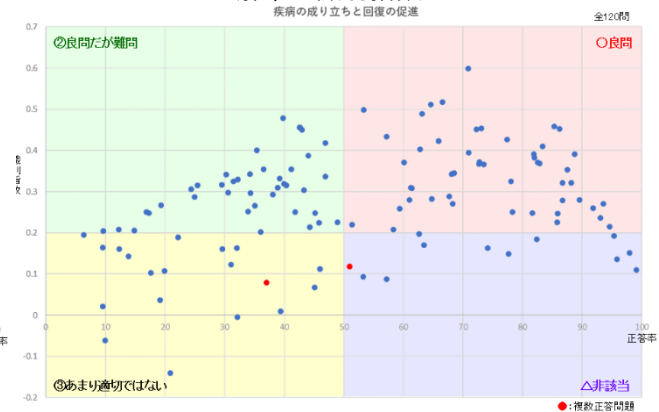


図5 基礎看護学 正解率－識別指数プロット

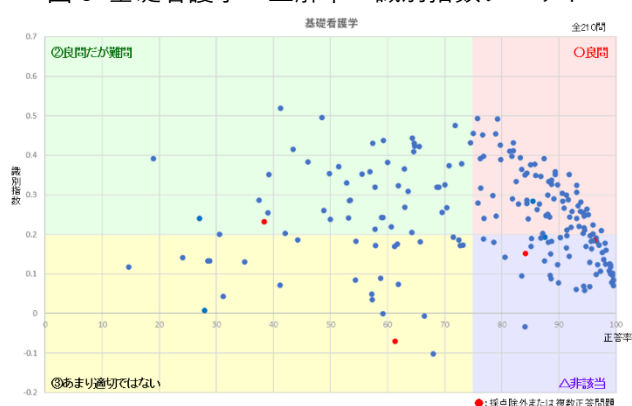
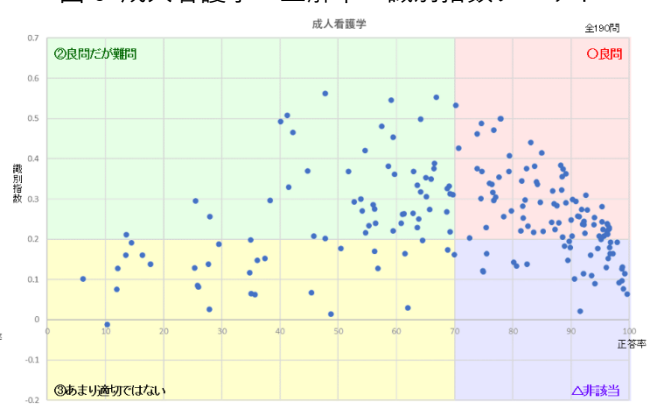


図6 成人看護学 正解率－識別指数プロット



次に、4科目のうち「人体の構造と機能」(120問)と「疾病の成り立ちと回復の促進」(120問)の2科目について、ルール抽出作業の第一段階である類似問題(質問テーマが同じ問題)の抽出を行った。

その結果、「人体の構造と機能」については、類似問題が極めて少なく、類似問題2組、該当4問であった。同様に「疾病の成り立ちと回復の促進」については、類似問題17組、該当38問であった。

2) 一般問題作問システムの開発

①作問システムの評価実験

実際の看護師国家試験を用いた評価実験を行った。

データセット

厚生労働省より提供を受けた看護師国家試験過去問題640問と、東京アカデミー(医療

系予備校)より提供を受けた模擬試験問題600問のうち300問を使用してデータセットを構築した。データセットを作成する上で、問題文に図表を含む問題、正解が複数存在する問題、数値を答える形式の問題、選択肢に組み合わせ形式を含む問題などを除外した。看護師国家試験の過去問題からは450問、模試試験問題からは268問をデータセットの対象とした。対象となるデータを分割し、表4に示す通り、訓練・検証・評価データセットをそれぞれ作成した。

表4 作成したデータセット数

	国家試験	模擬試験
	過去問題	問題
訓練	107	268
検証	107	0
評価	236	0

評価指標

誤答選択肢の評価には、人手で作成された実際の誤答選択肢に対する完全一致率 (R、P)、意味類似度 (Rs、Ps)、二部マッチングに基づく類似度 (Rcs、Pcs) を用いる。さらに生成した誤答選択肢の多様性を評価する多様性指標 (DV) の 7 指標を用いた。本研究で重視する DV は、全ての誤答選択肢のペアについて、コサイン距離を平均したものを評価値とする。DV は誤答選択肢同士の類似度が高いほど低い値となり、誤答選択肢同士がどれだけ多様かを示す。

比較手法

ベースラインの生成手法として、Zero-Shot、Five-Shot、SFT の 3 つを設定した。Zero-Shot では、システムプロンプトに生成要件を、ユーザープロンプトに問題文と正解選択肢を与えて生成させた。Five-Shot では、問題文、正解選択肢とその応答例を 5 つ与えて生成させた。SFT では、訓練データセットを用いてモデルをファインチューニングした後生成を行った。なお、ファインチューニングには QLoRA を採用した。すべての手法で Llama-3.1-Swallow-8B-Instruct-v0.5 を使い、ベースラインでは temperature は 0 とした。

実験結果

実験結果を表 5 に示す。提案手法においては temperature=(0.2,0.4,0.6) の組で生成させたものを Proposed_0.2-0.6、temperature=(0.8,1.0,1.2) の組で生成させたものを Proposed_0.8-1.2 と表中に示した。なお評価データセットにおける実際の誤答選択肢の DV の値は 0.133 であった。

実験結果より、提案手法は DV の点でベースライン 3 手法を大きく上回っている。よっ

て、ベースラインと比較して、提案手法はより多様な誤答選択肢を生成できたと言える。加えて、Proposed_0.8-1.2 において、DV は実際の試験問題と同程度の値となっており、生成した誤答選択肢が、実際の誤答選択肢と同程度の多様性を持つと言える。また誤答選択肢の品質を表す評価指標 Rs、Ps、Rcs、Pcs についても、多様性と品質にはトレードオフの関係があるが、提案手法はどちらも Five-Shot、SFT と比較して高い水準を保っている。Proposed_0.2-0.6 において、Rs、Ps、Rcs、Pcs は Zero-Shot と同程度の値となっており、多様性を大幅に向上しながら誤答選択肢の品質を維持できていると言える。

作成した誤答選択肢の例を図 7 に示す。図 7 において、Five-Shot で生成した誤答選択肢はすべて「左心室の内圧は」を含んでおり、誤答選択肢同士の多様性が失われてしまっている。一方、提案手法 Proposed_0.2-0.6 で作成した誤答選択肢は「左心室の内圧」以外の項目も含み、多様性が向上していることが確認できる。

図 7 作成した誤答選択肢の例

問題: 心周期に伴う心臓の変化で、収縮期の初期には心室の容積は変わらずに内圧が上昇している。このときの心臓で正しいのはどれか。
正解: 左心室の内圧は大動脈圧よりも低い。
Five-Shot で作成した誤答選択肢:
左心室の内圧は大動脈圧よりも高い。
左心室の内圧は収縮期を通して一定である。
左心室の内圧は拡張期に上昇する。
Proposed_0.2-0.6 で作成した誤答選択肢:
大動脈弁は開いている。
心室の壁は薄くなっている。
心室の内圧は一定である。

表 5 作問システムの評価実験結果

手法	<i>R</i>	<i>P</i>	<i>R_s</i>	<i>P_s</i>	<i>R_{cs}</i>	<i>P_{cs}</i>	<i>DV</i>
<i>SFT_0.0</i>	0.109	0.109	0.896	0.901	0.886	0.886	0.106
<i>Five-Shot_0.0</i>	0.097	0.097	0.896	0.898	0.884	0.884	0.118
<i>Zero-Shot_0.0</i>	0.102	0.102	0.892	0.895	0.881	0.881	0.109
<i>Proposed_0.2-0.6</i>	0.076	0.076	0.894	0.895	0.880	0.880	0.121
<i>Proposed_0.8-1.2</i>	0.060	0.060	0.889	0.887	0.874	0.874	0.134

2. 研究 1-2 校閲システム開発

今年度は先行研究で入手した教科書のテキストデータと、模擬試験問題、国家試験問題について出題基準および類似問題かどうかを判定しラベル付したデータを入手した。以降の作業は、次年度に取り組む計画である。

3. 研究 1-3 出題形式 (Taxonomy) 評価システム開発

本項目についても、「研究 1-2」と同様に今年度はラベル付したデータ入手までを行った。以降の作業は、次年度に取り組む計画である。

D. 考察

1. 研究 1-1 看護師等国家試験 一般問題の作問システム開発

1) 看護師国家試験（一般問題）過去問分析

先行研究では必修問題を対象としたため、良問の条件を正解率：90%以上 99%未満、識別指数：0.2 以上と定めることで大きな齟齬は生じなかった。しかし、一般問題は大項目、中項目が多岐にわたること、また Taxonomy も I'あるいは II 型も含まれるなど問題がより複雑で難易度が高いことから、必修問題に比べて正解率が低い傾向となっていると推察された。その結果、4 科目に対して一律の評価基準を設定することは適切ではないと考えられた。今年度行った 2 科目の一般問題の分析結果より、一般問題は必修問題と比べてルール抽出が難しいことが考えられた。今後、基礎看護学、成人看護学について

も類似問題の分析を行い、ルール抽出が可能であれば、ルールブックの作成を試みるとともに作問システムのプロンプトとして活用する。

2) 一般問題作問システムの開発

提案手法は DV の点でベースライン手法を大きく上回っており、多様性の向上が確認された。一方で、誤答選択肢の品質を表す評価指標においては、多様性とのトレードオフの関係が見られるが、提案手法では高い水準を維持していた。Five-Shot で生成した誤答選択肢は過剰に類似した表現を含む傾向があり

(図 7)、誤答選択肢同士の多様性が失われている。一方、提案手法で作成した誤答選択肢は異なる観点を含み、多様性が向上していることが確認できた。

Temperature を上げることで誤答選択肢同士の多様性を高めることができるが、その一方で実際の誤答選択肢から離れた内容が生成される傾向がある。提案手法では LLM-as-a-Judge による排除・抽出機構を導入することで、この問題を抑制しつつ多様性を確保できることが示された。

2. 【研究2】ICT・AI 技術を活用した看護師等国家試験問題作問システムの実装と評価

担当：(研究分担者) 米倉、宇佐美、佐伯、佐々木、佐居、(研究協力者) 伊藤、西崎、宮本、鈴木、野崎、水戸、木村、山本

B. 研究方法

1. 研究 2-1 ルールブックならびに作問システムの試験委員会での利用可能性調査

本調査では、看護師国家試験委員を対象として、先行研究で作成した作問ルールブックを作問時に使用してもらい、ルールブックの受容性、すなわちルールブックの活用による作問負担感への影響や作問における有用性を評価した。

「作問ルールブック」とは、先行研究において、看護師国家試験 10 年分（第 102 回～

第 111 回）の必修問題 500 問について、改善提案と作問例を大項目ごとにまとめたものであり、試験委員が閲覧可能な Excel ファイルとして提供した（表 6）。調査協力者には Google フォームで作成した web 質問紙にて質問への回答を依頼した。

質問内容は、「ルールブックは作問の参考になったか」、「ブレインストーミングに役立ったか」、「ルールブックを使用しない時と比べて作問作業の負担感はどうだったか」、「ルールブックは自身の自由な発想を阻害したか」、「ルールブックの改善点や意見」であった。

調査実施にあたり、聖路加国際大学研究倫理審査委員会の承認を得た（承認番号：25 - A034）。

表 6 作問ルールブックの形式

出題基準		過去問題データ					選択肢のタイプ	問のタイプ	評価	コメント	改善提案	作問例
中項目	小項目	出題年	問題番号	問題文	選択肢 (1) (2) (3) (4) (5)	正答						
							名詞句 数値 文 図式	what where how why	○ ① ② ③ △	・類似問題との比較 ・解答者数の分布傾向等		

2. 研究 2-2 国内外の他資格・分野の国家試験等の ICT・AI 活用状況並びに機密性の担保に関する調査

国内で先行する調査の報告書(厚生労働行政推進調査事業費（地域医療基盤開発推進研究事業）「保健師助産師看護師国家試験におけるコンピュータの活用に関する研究」令和 4 年度総括・分担研究報告書(宮本 他, 2023)⁷⁾; 以

下「報告書」と表記)を参照し、情報を整理した。整理した情報をもとに、今後の調査や議論の方向性を研究班メンバーで検討した。

3. 研究 2-3 一般問題の作問システムによる問題評価とテストシステムの検討

日本看護学校協議会の加盟校並びに看護系大学（JANPU 会員校）を対象に本研究の主旨を説明し、令和 8 年度、9 年度に行う一般

問題の問題評価調査（TAO を用いた看護師一般問題模擬試験形式）に協力の意向を示す学校（約 30 校を目標とする）の代表教員によるプロジェクトチームを形成する方向で、検討を進めた。

一方で、一般に行われている看護師国家試験模試に本研究班で作成した問題を使用することで、問題の質評価を行うことが可能か、模擬試験を行う予備校、出版社との協働も視野に入れて計画を検討した。

C. 研究結果

1. 研究 2-1 ルールブックならびに作問システムの試験委員会での利用可能性調査

令和 7 年度中の回答者は 1 名であった。対象の背景は、看護師臨床経験 7 年、看護教員経験年数 15 年、看護師国家試験委員の経験年数 3 年であり、現在は管理職業務に就いていた。回答内容の概要を表 7 に示す。

表 7 作問ルールブック利用可能性評価の調査回答（概要）

問 1：作問ルールブックは作問の参考になりましたか
「1：参考になった」
問題の問いの立て方、肢の並びや質によって難易度が変わることがわかり、作問をする上でのヒントとなった。さらにこのような作問であると良問となる、といったヒントも有効に活用できた。また、すでに作問した内容についても難易度やその理由が明記されていたため、作問をするうえで留意しながら取り組むことができた。
問 2：作問ルールブックは、ブレインストーミングに役立ちましたか。
「1：役立った」
会議の議論の中で指摘を受けた点について、グループ内のディスカッションが行き詰った時に、ルールブックを読み解きながらディスカッションを発展することができた。ル

ールブックがない状態の時には、各々が考えている難易度などの方向性がうまく擦り合わない時もあったが、ルールブックを介してディスカッションすることで効果的な意見交換につなげることができた。

問 3：作問ルールブックを使用した作問作業は、使用しない時と比べて負担感はどのくらいでしたか。

「2：どちらかといえば負担感は小さかった」

これまでは、自分の持ちうる知識・経験値・出題基準をもとに頭の中でアイディアを捻出することに大変負担を感じていた。必修としての難易度を考えながら作問する過程は、自身の思考だけで十分に整えることができず、既出の内容に寄った問題しか捻出できないなど苦慮することが多かった。しかし、ルールブックを活用することで、アイディアの方向性が導きやすくなり、負担感は減ったように感じる。一方、今回のルールブックの情報量では打開できない部分もあったため、「2」（どちらかといえば・・・）の評価とした。今後の開発に期待を寄せている。

問 4：作問ルールブックは、ご自身の自由な発想を阻害しましたか。

「1：阻害しなかった」

ルールブックが指し示しているのは、あくまで方向性として理解しているので、自由な発想の阻害にはならなかった。逆に、ルールブックのヒントによって、この方向性の問題でいいのか？という疑問が解消されたり、別の角度で考えてみることのヒントになったりしたので、発想を広げるのにも役立った。

問 5：今回利用していただいたルールブックについて、改善点やご意見等

ルールブックの情報量が増えて、ヒントとなる視点が発展されることを願っている。

2. 研究 2-2 国内外の他資格・分野の国家試験等の ICT・AI 活用状況並びに機密性の担保に関する調査

報告書では、日本の他資格・分野の試験の実態、諸外国での看護師等免許取得試験についてまとめられていた。看護職の免許取得試験において CBT が活用されているのは、アメリカ・カナダ(NCLEX)、イギリスであること、韓国においては CBT 導入が検討されていることが確認された。また、国内では他資格の国家資格・国家試験で CBT が活用されている事例は認められたが、限定的であることが確認された(宮本 他, 2023)。

3. 研究 2-3 一般問題の作問システムによる問題評価とテストシステムの検討

今年度はプロジェクトチームの形成には至らなかったが、国家試験模試を行っている業者(予備校、出版社)複数と連携をとり、今後の調査に向けて協働の可能性を探ることとなった。

D. 考察

1. 研究 2-1 ルールブックならびに作問システムの試験委員会での利用可能性調査

アンケートの結果から、ルールブックを活用することで、作問のヒントが得られたこと、また国家試験作問会議においてもグループ内の検討に有効であったことが明らかとなった。また自身の自由な発想を阻害することなく、アイディアの方向性の示唆が得られ負担感が減ったとの回答から、ルールブックは実際の作問過程において有効に活用される可能性が示唆された。一方で、より多くの情報が求められていることから、定期的なブラッシュアップと情報量の更新が求められていると考えられる。

今回は 1 名の回答であったことから、次年度も継続的に調査を行い、より広く意見を収集することで、ルールブックの利用可能性についてさらに検討する事が必要である。

2. 研究 2-2 国内外の他資格・分野の国家試験等の ICT・AI 活用状況並びに機密性の担保に関する調査

先行研究(宮本 他, 2023)より、国内他資格、諸外国の看護師資格試験における CBT の活用状況に関する、制度・仕組みの全体像は十分把握できていることが確認できた。CBT を活用した試験の基本構造は短期間で大きく変わるものではなく、本課題で今後新規に調査をする必要性は限定的であると考えられた。また、作問プロセスや問題の機密性保持、試験委員体制等は機密情報であるため、先行調査では明示されておらず、情報収集事態が困難であると考えられた。

今後の本課題における調査、検討の方向性として、(1)諸外国における看護師資格試験について、先行する調査以降の変更を調査すること、(2)(1)を踏まえ、すでに海外で CBT で実施されている方式を日本に適用した場合の検討事項、実現可能性を具体的にシミュレーションすること、の 2 点が挙げられた。

(1)については、アメリカ等で実施されている NCLEX は 2023 年以降に Next Generation NCLEX(NCLEX-NGN)として試験問題の傾向が変わったため、具体的な変更点を調査し整理すること、韓国における看護師国家試験の CBT への移行状況を明らかにすることが課題となった。

(2)については、諸外国での CBT による看護師資格試験の実施に関する具体的な検討事項(受験資格の要件、付与時期、試験回数・スケジュール、試験問題のプール方法、試験問題の形式、等)を整理し、現行の日本の看護師国家試験からの移行における課題と実現可能性を整理し、日本において実現可能な CBT による看護師国家試験のシナリオを提示することを今後の研究課題とすることとした。

E. 結論

1. 【研究 1】 ICT・AI 技術を活用した看護師等国家試験問題作問システムの構築

令和 7 年度は看護師等国家試験の一般問題出題数の上位 4 科目にあたる、人体の構造と機能、疾病の成り立ちと回復の促進、基礎看護学、成人看護学に焦点を当て、作問システムの開発に向けた作業を開始した。

当研究班の先行研究と同様に、本研究においても誤答肢生成のための作問システムを開発することとした。開発のための基礎データとして、厚生労働省から国試過去問 10 年分（人体の構造と機能 120 問、疾病の成り立ちと回復の促進 120 問、基礎看護学 210 問、成人看護学 190 問）計 640 問、看護師国家試験模擬試験同 4 科目、各 150 問ずつ合計 600 問を入手した。

国試過去問 640 問について、問題の評価（良問：評価○、良問だが難問：評価②、あまり適切ではない問題：評価③、非該当（いずれにも該当しない）：評価△）を決定するため、良問とその他を判別する正解率、識別指数の閾値を決定した。正解率については過去 10 年間の平均値を算出し、これに基づき閾値を決定した。識別指数は、先行研究と同様に 0.2 を閾値とした。

各科目の正解率の平均値からそれぞれ正解率の閾値を決定（各 50%、50%、75%、70%）し、正解率（横軸）、識別指数（縦軸）をプロットし、問題の評価を図示した。

その結果、各科目の良問数は、「人体の構造と機能」58 問（48.7%）、「疾病の成り立ちと回復の促進」49 問（40.8%）、「基礎看護学」76 問（36.5%）、「成人看護学」78 問（41.1%）であった。

次に、4 科目のうち「人体の構造と機能」（除外問題 1 問を除いた 119 問）と「疾病の成り立ちと回復の促進」（120 問）の 2 科目について、類似問題（質問テーマが同じ問題）の抽出を行った。その結果、「人体の構造と機能」に

ついては、類似問題が極めて少なく、類似問題 2 組、該当 4 問、「疾病の成り立ちと回復の促進」については、類似問題 17 組、該当 38 問であった。今年度行った 2 科目の分析結果より、類似問題が少なく必修問題のようなルール抽出は難しいことが推察された。今後、基礎看護学、成人看護学についても類似問題の分析を行い、ルール抽出が可能であれば、ルールブックの作成を試みるとともに作問システムのプロンプトとして活用する。

作問システム開発では、看護師国家試験の一般問題を題材として、誤答選択肢同士の多様性の問題に着目した。提案手法では、生成結果の多様性を高めるために temperature（出力の多様性レベルを示すパラメータ）を高い値に設定し、複数の temperature で得られた候補を集約した。加えて、非妥当な誤答選択肢を除外し、適切な誤答選択肢を抽出する目的で LLM-as-a-Judge を用いた自動判定を導入した。

提案手法では、生成機構、排除機構、抽出機構の三つの機構を用いて誤答選択肢の生成を行った。生成機構では、誤答選択肢を Five-Shot、ファインチューニング（SFT）の二つの手法でそれぞれ実際の誤答選択肢の数だけ生成させた。生成して得られたすべての誤答選択肢を集約し、これを次の排除機構に与えた。排除機構では、生成機構より与えられた誤答選択肢それぞれについて、問題文に対する正解になってしまう可能性があるか LLM を用いて判定し、正解になりうると判定された誤答選択肢を排除した。抽出機構では、与えられた誤答選択肢から、LLM が適切と判断したものを実際の誤答選択肢の数だけ抽出した。適切な誤答選択肢の基準はシステムプロンプトを通して与えるものとした。

その結果、本手法により、実際の試験問題と同程度の多様性を持つ誤答選択肢を、高い品質を維持しながら自動作成することに成功した。

本研究により作成した誤答選択肢が、実際の作問現場でどれほど活用可能かについては次

年度詳細に検証する計画である。

2. 【研究2】ICT・AI 技術を活用した看護師等国家試験問題作問システムの実装と評価

令和7年度は、先行研究で作成した必修問題作問ルールブックを、国家試験作問時に試験委員が使用することで、作問の負担感が軽減するか、ルールブックの受容性、有用性について評価した。作問ルールブックは、試験委員が閲覧可能なExcelファイルで提供した。その結果、ルールブックの活用により、作問のヒントが得られ、作問会議に於いてもグループ内の検討に有効であったことが明らかとなった。また作問者の自由な発想を阻害せず、アイディアの方向性の示唆が得られ負担感が減ったとの回答が得られ、実際の作問過程において有効に活用される可能性が示唆された。一方で、ルールブックの記載内容より多くの情報も求められ、定期的なブラッシュアップと情報の更新が必須であると考えられた。次年度も継続的に調査を行いより広く意見を収集することで、ルールブックの利用可能性についてさらに検討する事が必要である。

また国内外の他資格・分野の国家試験等のICT・AI活用状況並びに機密性の担保について、先行研究(宮本 他, 2023)より、国内他資格、諸外国の看護師資格試験におけるCBTの活用状況に関する、制度・仕組みの全体像は十分把握できていることが確認できた。今後の本課題における調査、検討の方向性として、(1)諸外国における看護師資格試験について、先行する調査以降の変更を調査すること、(2)(1)を踏まえ、すでにCBTで実施されている方式を日本に適用するとした場合の検討事項、実現可能性を具体的にシミュレーションすること、の2点が挙げられた。

F. 健康危険情報

該当なし

G. 研究発表

1. 論文発表

伊藤 晃、荒瀬由紀、「看護師国家試験一般問題の誤答選択肢自動生成手法の検討」、言語処理学会第32回年次大会発表論文集、pp.3371-3376、2026年3月。

2. 学会発表

林 直子、米倉佑貴、木村理加、佐居由美、三浦友理子、宮本千津子、鈴木久美、佐々木幾美、佐伯由香、「看護師国家試験必修問題の作問における大規模言語モデル(LLM)活用の可能性-看護教員による質的評価」、第45回日本看護科学学会学術集会、新潟、2025年12月。

木村理加、米倉佑貴、林 直子、佐居由美、三浦友理子、宮本千津子、佐々木幾美、佐伯由香、「大規模言語モデルによる支援を受けて作成した看護師等国家試験問題の質の評価」、第45回日本看護科学学会学術集会、新潟、2025年12月。

米倉佑貴、林 直子、木村理加、佐居由美、三浦友理子、宮本千津子、佐々木幾美、佐伯由香、「看護師等国家試験模擬試験のComputer Based Testing(CBT)形式での受験経験の評価」、第45回日本看護科学学会学術集会、新潟、2025年12月。

林 直子、木村理加、佐居由美、三浦友理子、米倉佑貴、山本真実、宮本千津子、佐伯由香、佐々木幾美、鈴木久美、「交流集会：看護師等国家試験問題作問におけるICT・AI活用の可能性—大規模言語モデル(LLM)活用の試み」、第45回日本看護科学学会学術集会、新潟、2025年12月。

伊藤 晃、荒瀬由紀、「看護師国家試験一般問題

の誤答選択肢自動生成手法の検討」、言語処理学会第 32 回年次大会(NLP2026)、宇都宮、2026 年 3 月。

H. 知的財産権の出願・登録状況

該当なし

引用文献

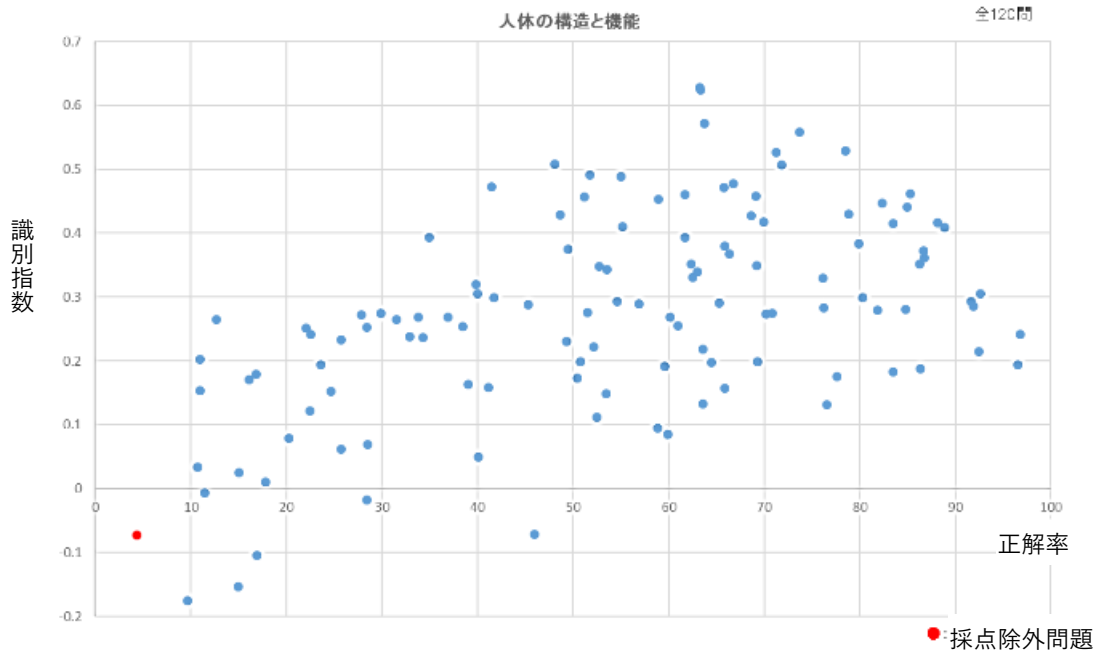
- 1) 林直子、徳永健伸、宇佐美慧、佐伯由香、佐々木幾美、米倉佑貴、佐居由美、厚生労働科学研究費補助金（政策科学総合研究事業（臨床研究等 ICT 基盤構築・人工知能実装研究事業））、保健師助産師看護師国家試験の問題作成の支援と効率化に向けた ICT・AI 技術等の活用策の検討のための研究、令和 4 年度～6 年度総合研究報告書、https://mhlw-grants.niph.go.jp/system/files/download_pdf/2024/202403001B.pdf
- 2) Yusei Kido, Hiroaki Yamada, Takenobu Tokunaga, Rika Kimura, Yuriko Miura, Yumi Sakyō and Naoko Hayashi, Automatic Question Generation for the Japanese National Nursing Examination Using Large Language Models, Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024), pp.821-829, 2024.
DOI:10.5220/0012729200003693.
- 3) 城戸祐世、山田寛章、徳永健伸、木村理加、三浦友理子、佐居由美、林直子、言語モデルを用いた看護師国家試験問題の誤答選択肢自動生成、言語処理学会第 31 回年次大会発表論文集、pp.1074-1079, 2025
- 4) Yusei Kido, Hiroaki Yamada, Takenobu Tokunaga, Rika Kimura, Yuriko Miura, Yumi Sakyō and Naoko Hayashi, Evaluation of LLM-Generated Distractors of Multiple-Choice Questions for the Japanese National Nursing Examination, Proceedings of the 17th International Conference on Computer Supported Education(CSEDU 2025), pp.754-764, 2025.
DOI:10.5220/0013460300003932.
- 5) Rajiv Jhangiani. Guidelines for writing effective distractors for multiple-choice questions, (2026.5.28 参照).
<https://thatpsychprof.com/wp-content/uploads/2016/12/Guidelines-for-writing-effective-distractors-for-multiple-choice-questions.pdf>.
- 6) 伊藤 晃、荒瀬由紀、「看護師国家試験一般問題の誤答選択肢自動生成手法の検討」、言語処理学会第 32 回年次大会発表論文集、pp.3371-3376、2026 年 3 月。
- 7) 宮本千津子、小黒道子、勝山貴美子、西崎祐史、鋪野紀好、田所由利子、窪田和巳（2023）。保健師助産師看護師国家試験におけるコンピュータの活用に関する研究：令和 4 年度総括・分担研究報告書（課題番号 22IA2008）。厚生労働科学研究費補助金健康安全確保総合研究分野地域医療基盤開発推進研究事業。厚生労働科学研究成果データベース（文献番号 202222070A）。
<https://mhlw-grants.niph.go.jp/project/164987>

資料

看護師国家試験問題（第 105 回～第 114 回） 正解率－識別指数プロット

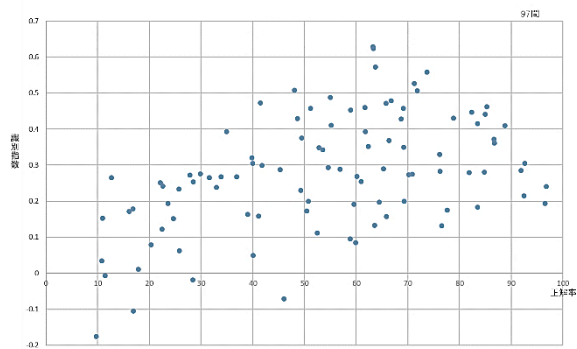
1. 人体の構造と機能

(1) 全 120 問

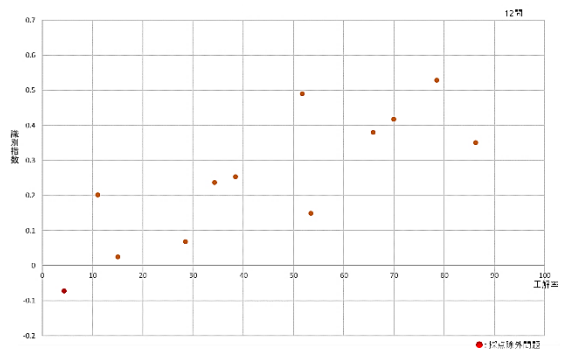


(2) Taxonomy 別

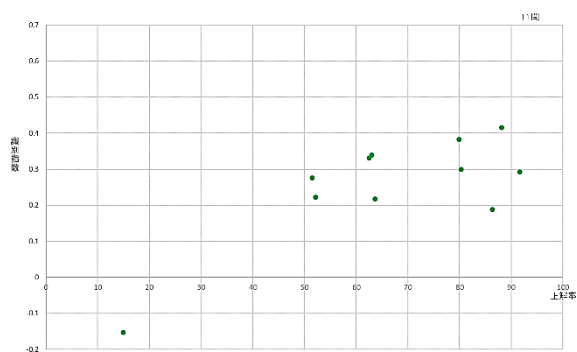
Taxonomy I 単純想起型



Taxonomy I' 推定型



Taxonomy II 解釈型



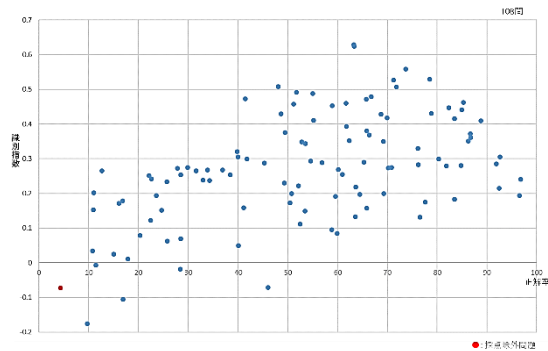
(3) 目標別

目標Ⅰ．正常な人体の構造と機能について基本的な理解を問う。

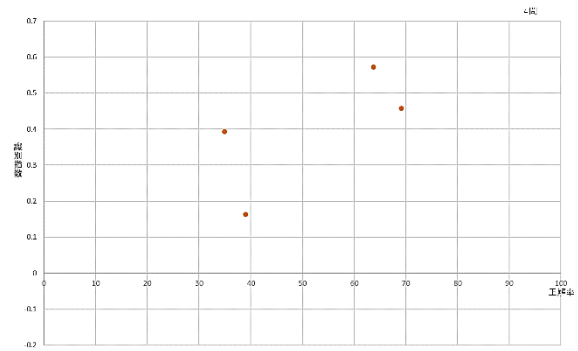
目標Ⅱ．フィジカルアセスメントおよび日常生活の営みを支える看護に必要な人体の構造と機能について基本的な理解を問う

目標Ⅲ．疾病の成り立ちとの関連において、人体の構造と機能について基本的な理解を問う。

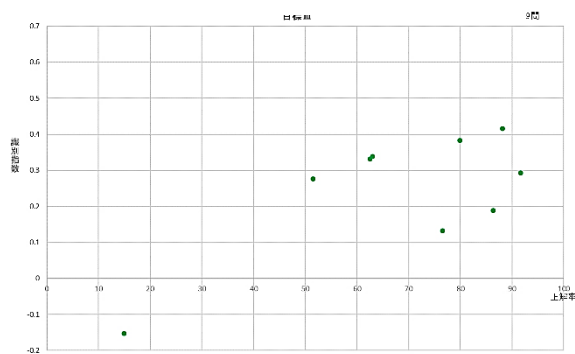
目標Ⅰ



目標Ⅱ

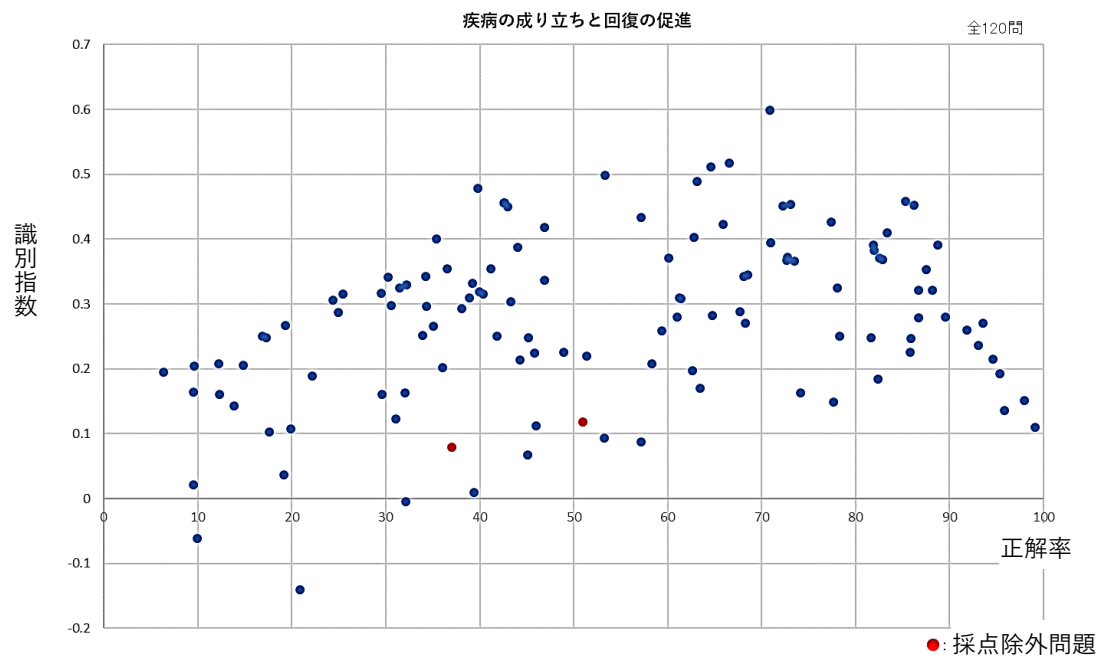


目標Ⅲ



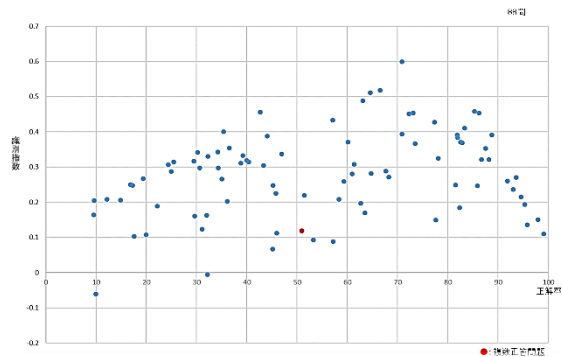
2. 疾病の成り立ちと回復の促進

(1) 全 120 問

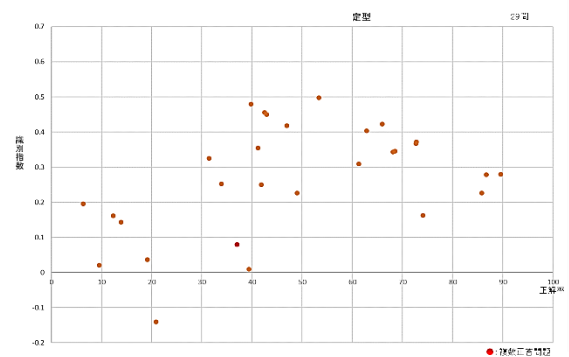


(2) Taxonomy 別

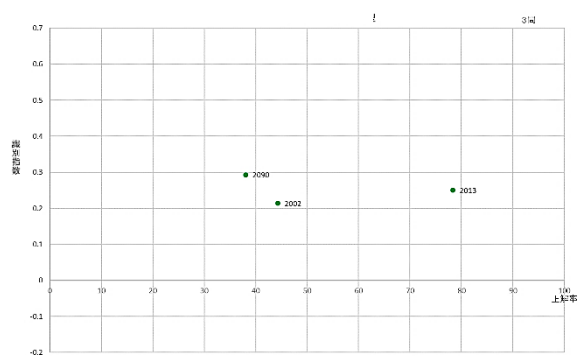
Taxonomy I 単純想起型



Taxonomy I' 推定型



Taxonomy II 解釈型



(3) 目標別

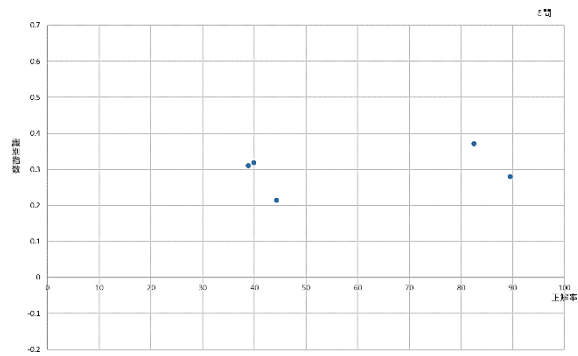
目標Ⅰ．健康から疾病を経て回復に至る過程について基本的な理解を問う。

目標Ⅱ．疾病の要因と生体反応について基本的な理解を問う。

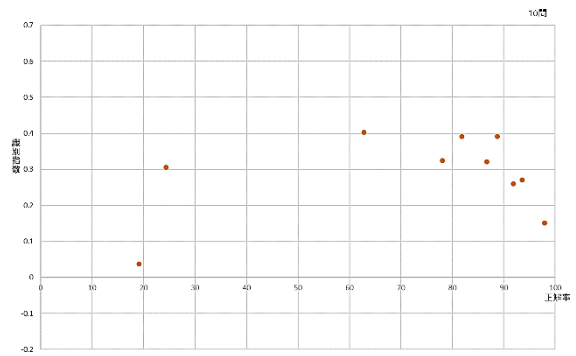
目標Ⅲ．疾病に対する診断・治療について基本的な理解を問う。

目標Ⅳ．各疾患の病態と診断・治療について基本的な理解を問う。

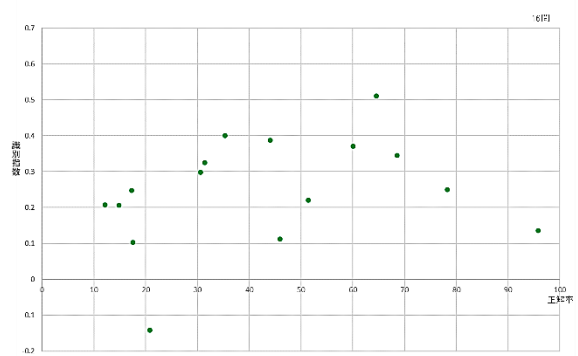
目標Ⅰ



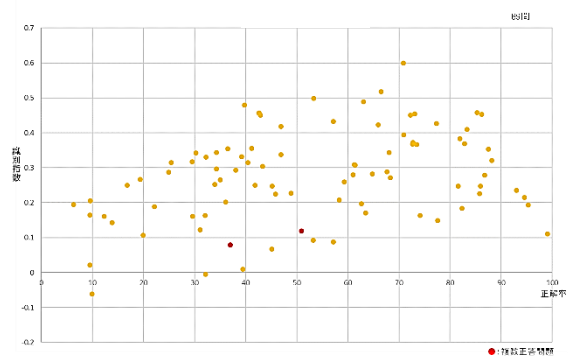
目標Ⅱ



目標Ⅲ

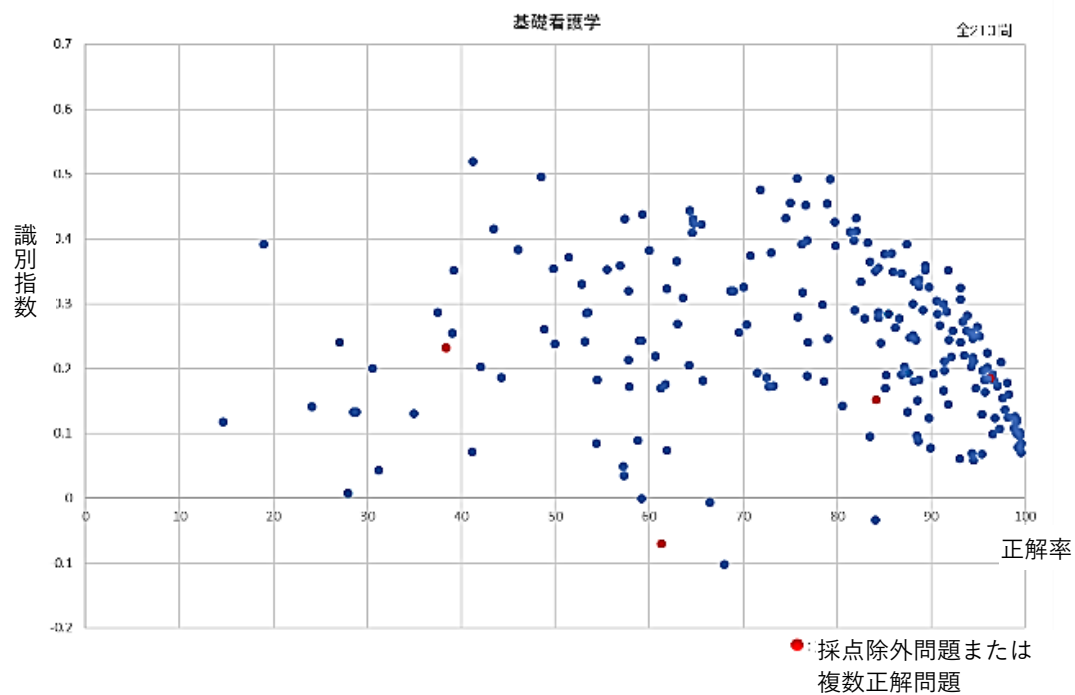


目標Ⅳ



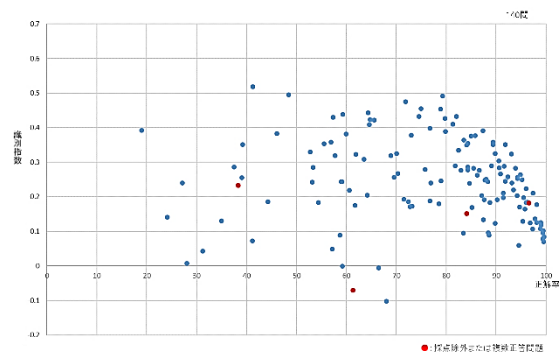
3.基礎看護学

(1) 全 210 問

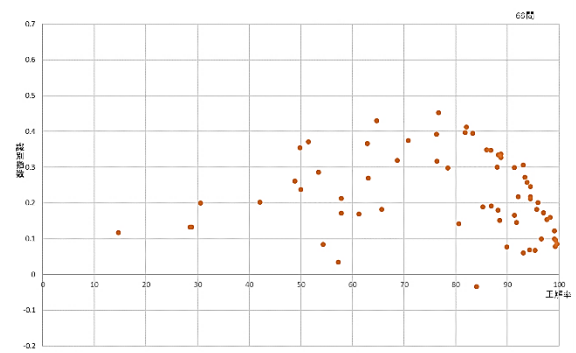


(2) Taxonomy 別

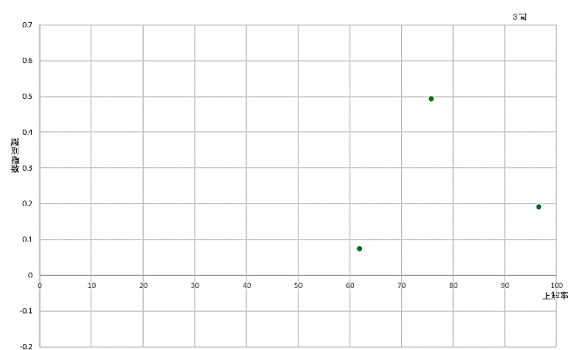
Taxonomy I 単純想起型



Taxonomy I' 推定型



Taxonomy II 解釈型



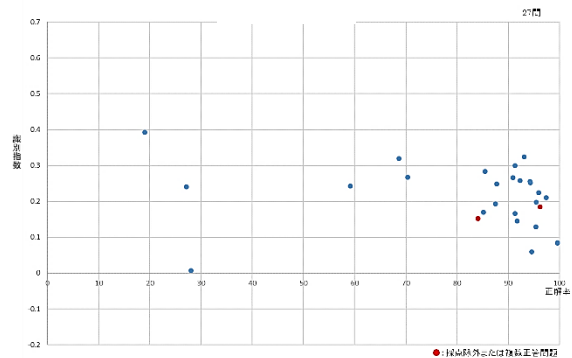
(3) 目標別

目標Ⅰ．看護の概念及び展開について基本的な理解を問う。

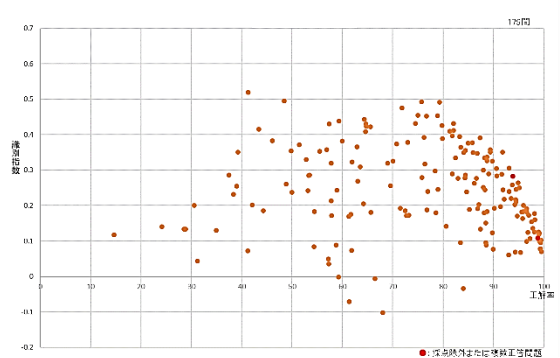
目標Ⅱ．基礎的な看護技術と適用のための判断プロセスについて基本的な理解を問う。

目標Ⅲ．保健・医療・福祉の中で看護の果たす役割について基本的な理解を問う。

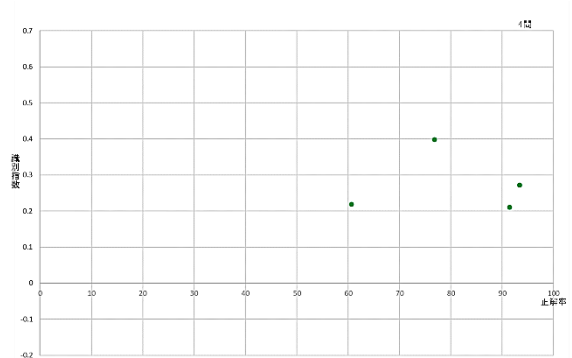
目標Ⅰ



目標Ⅱ

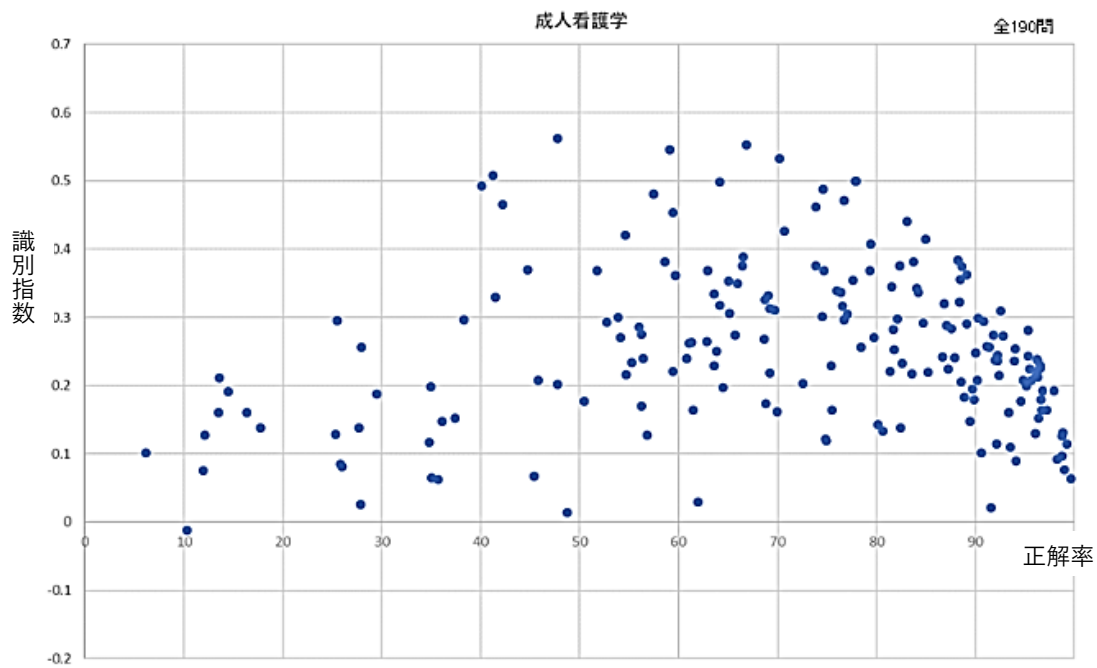


目標Ⅲ



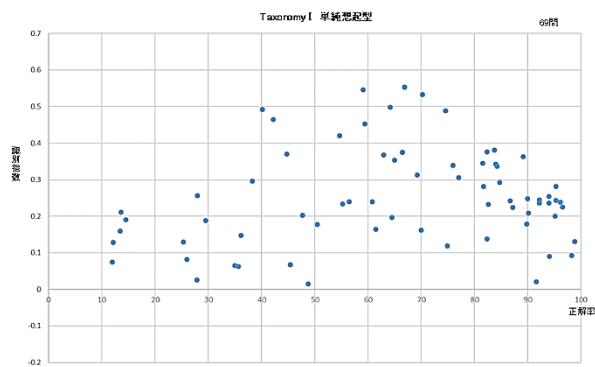
4. 成人看護学

(1) 全 190 問

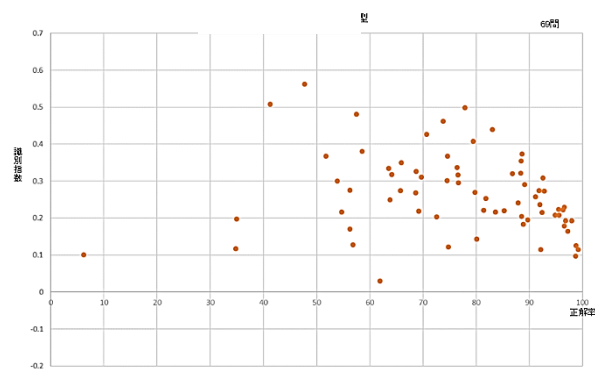


(2) Taxonomy 別

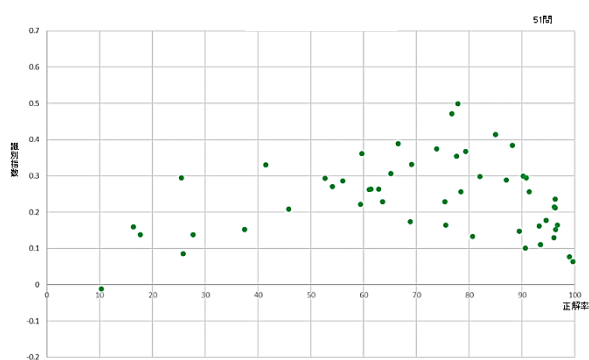
Taxonomy I 単純想起型



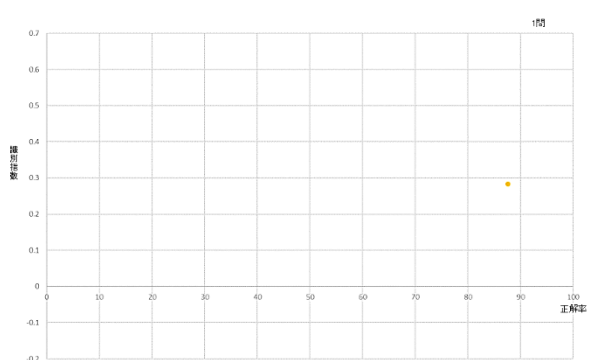
Taxonomy I' 推定型



Taxonomy II 解釈型



Taxonomy III 問題解決型



(3) 目標別

目標Ⅰ．成人各期の健康保持・増進や疾病の予防について基本的な理解を問う。

目標Ⅱ．急性期にある患者と家族の特徴を理解し看護を展開するための基本的な理解を問う。

目標Ⅲ．慢性疾患がある患者と家族の特徴を理解し看護を展開するための基本的な理解を問う。

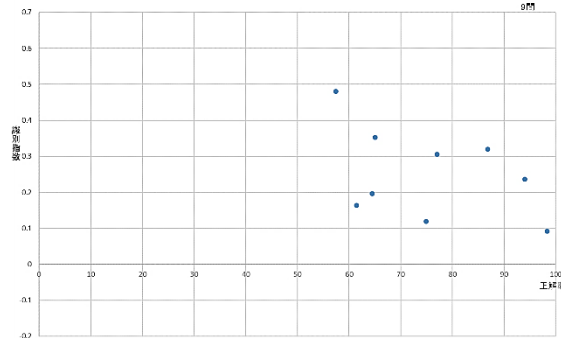
目標Ⅳ．リハビリテーションの特徴を理解し看護を展開するための基本的な理解を問う。

目標Ⅴ．がん患者と家族の特徴を理解し看護を展開するための基本的な理解を問う。

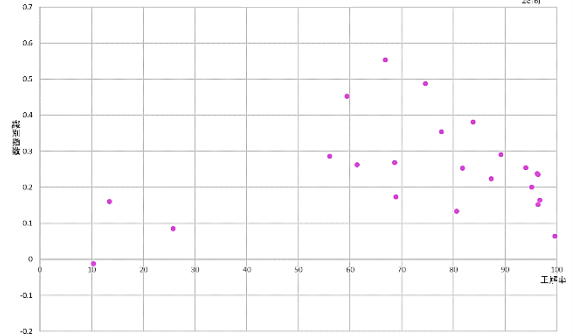
目標Ⅵ．終末期にある患者、および緩和ケアを必要とする患者と家族の特徴を理解し看護を展開するための基本的な理解を問う。

目標Ⅶ．各機能障害のある患者の特徴および病期や障害に応じた看護について基本的な理解を問う。

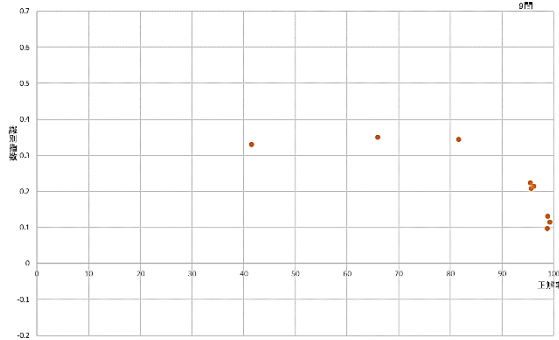
目標Ⅰ



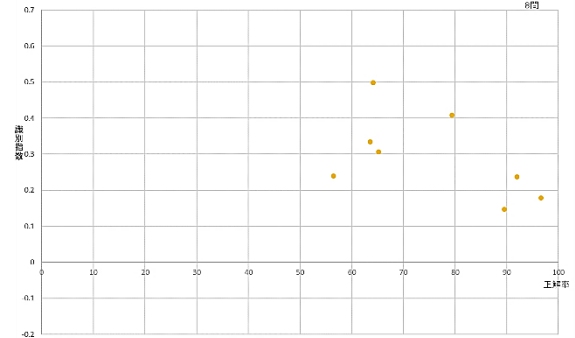
目標Ⅱ



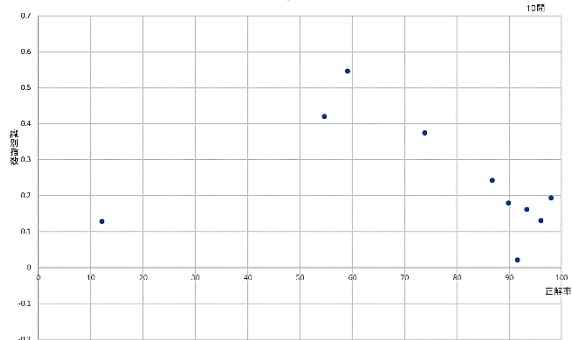
目標Ⅲ



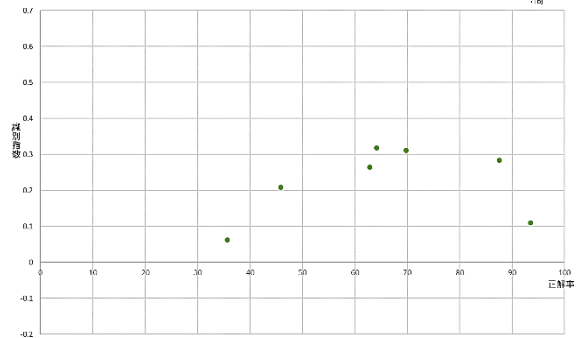
目標Ⅳ



目標Ⅴ



目標Ⅵ



目標Ⅶ

