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

研究代表者　林　直子

令和7（２０２５）年 5月

目　　次

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））

総合研究報告書

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

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

本研究は、保健師助産師看護師国家試験（以下「看護師等国家試験」）において、ICT・AI 技術等を活用した具体的な作問システムを検討すること(研究 1)、また看護師等国家試験におけるコンピュータを活用した試験（CBT：Computer-based Testing）の実装に向けて、CBT システムの試用版を開発・試行し、試験問題の妥当性について正答率、識別指数、IRT スコア等から評価すると共に、出題形式、問題管理システム、受験者側の受容性に関する調査を行い CBT 導入に向けた課題を明示すること(研究 2)を目的とした。研究は 3 年計画で遂行した。

【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

初年度(令和 4 年度)は、過去 10 年間の看護師国家試験必修問題（500 問）を、一定の基準をもとに 5 区分の評価（良問、易しすぎる問題、良問だが難問、あまり適切でない問題、いずれにも該当しない問題）に分類し、問題形式、解答形式の観点で小項目ごとの問題の分析を行い、小項目単位での良問ルールの抽出を開始した。また、既存の大規模事前学習済み言語モデルを試用して作問を行い、今後の作問システム開発への示唆を得た。

2 年目(令和 5 年度)は、看護師国試過去問題（必修問題）の分析結果を統合した改善提案（ルールブック）の作成と、3 種類の大規模言語モデル（ChatGPT 3.5、ChatGPT 4、JSLM）に基づく 4 タイプの作問システム（gpt4、gpt3.5-few shot learning、gpt3.5-Fine Tuning、JSLM）を用いて試験的作問を行い、専門家による問題評価と各モデルの生成性能を評価するための自動評価指標の作成を試みた。過去問題の分析の結果 258 の改善提案が示され、提案内容の分析の結果、200 の小項目に対して 147 のルールが抽出された。また大規模言語モデルの活用可能性の検討の結果、gpt3.5-FT が他と比べて再現性、精度の評価指標値が良いことが示された。

最終年度（令和 6 年度)は、6 種類・7 モデルの大規模言語モデル（LLM）を用いて看護師国試必修問題作成のうち、誤答肢作成を支援するシステムの作成を試みた。看護師国試委員経験者 15 人を対象として、大規模言語モデルにより作成した誤答肢案を提示して誤答肢を作成する群(AI アシスト有り群)と誤答肢案を提示せずに従来通りの方法で作成する群(AI アシスト無し群)の 2 群に分け、国試問題 10 問の誤答肢の作成状況について比較した。その結果、LLM を用いて作成された誤答肢案 102 のうち 48（47%）が作問者（対象者）により採用され、さらに採用はされなかったが妥当と判断されたものは 22 あったことから LLM が生成した誤答肢案の 69%（70/102）

が妥当だと判断された。また採用された 48 の誤答肢案のうち、11（22%）は誤答肢案を参照しないで作問するグループ（AI アシスト無し群）が作成した回答と一致したことから、AI アシストによる誤答肢案の一部は看護師国試作問者の思考と同程度で誤答肢を生成していると考えられ、その質の高さが確認された。また、AI アシスト有り群は AI アシスト無し群よりも作問時間が短く、作問の負担が軽く、LLM 生成の誤答肢案は参考になった、ブレーンストーミングに役立った、自由な発想は阻害されなかったという回答が得られた。これらのことから LLM による国試問題の誤答肢作成支援が可能であることが示唆された。次に、大規模言語モデル（LLM）を利用して作成した誤答肢案を参照して作問された問題（AI アシスト有り問題）と誤答肢案なしで作問された問題（AI アシスト無し問題）の正答率、識別指数、並びに問題の質を評価することを目的として、国内の大学、専門学校の看護学生 384 人を対象に模擬試験形式の調査を、また看護教員 46 人を対象に、各問題の誤答肢の質評価のための質問紙調査を行った。学生対象の模擬試験形式の調査では、AI 支援の有無による得点率に大きな差は見られず、LLM を活用した誤答肢作成が教育的に実用可能であることが示された。教員を対象とした調査では、AI 支援によって作成された問題の多くが従来の作問と同等の質であると評価された。特に「薬理作用」や「看護活動の場と機能」など一部の分野においては、LLM の支援が有効であることが確認された。一方、「社会保障」や「看護技術」に関しては、LLM の適用に限界がある可能性も指摘された。以上の結果から、LLM は看護師国家試験の必修問題の誤答肢作成における支援ツールとして有用であることが明らかとなった。

【研究 2】看護師等国家試験への CBT 実装の可能性の検討

初年度は、CBT システムの試験運用に向けた問題収集（500 問）と動画や音声等のマルチメディアを活用した問題の作成、さらに CBT 実装可能性の検討に向けたインタビュー調査の準備を行った。

2 年目は、看護系大学と看護専門学校計 3 校の学生 27 名を対象に、10 問のサンプル問題について CBT システムと筆記試験の双方での解答を依頼し、さらに CBT システム導入の可能性に関するフォーカスグループインタビュー調査を実施した。また看護教員 17 名を対象に、看護師国家試験への CBT 導入の可能性に関するインタビュー調査を行った。その結果、学生、教員ともに看護師国試への CBT 導入の受容性が示されたが、一方で CBT の実施に向けた教員側の意識改革などの課題も明らかとなった。

最終年度は、筆記試験と CBT の解答形式の違いによる正答率や解答のしやすさの相違を明らかにすることを目的として、国内の看護系大学、看護専門学校の学生 384 人を対象に模擬試験形式の調査を行った。筆記試験、CBT の試験形式間の比較では、両群間で試験の正答率には差がみられなかったものの、疲労感は CBT 群の方が有意に低く（$p=0.005$）、解答に集中できたか（$p=0.001$）、文字の読みやすさ（$p=0.002$）、時間配分のしやすさ、消去法のしやすさ、メモの取りやすさなど、解答のしやすさについては、全ての項目において筆記試験群の方が肯定的な評価をした者の割合が有意に高かった（$p<0.001$）。今後、看護師等国家試験を CBT で実施することを検討する際には、消去法やメモ等を筆記試験での解答と同様に実行できる機能を使用可能とするとともに、受験者がシステムに習熟する機会を確保し、試験内容とは関係のない要素が試験結果に影響しないよう、十分に準備することが必要であることが示唆された。

研究分担者氏名　所属　　職位
徳永 健伸　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
乾 健太郎　　東北大学大学院　教授
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
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
山田 寛章　　東京科学大学　助教
城戸 祐世　　東京科学大学　修士課程学生
三浦 友理子　聖路加国際大学　准教授
木村 理加　　聖路加国際大学　助教

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

担当：(研究代表者)林、(研究分担者)徳永、佐伯、佐々木、佐居、(研究協力者)乾、宮本、鈴木、山田、城戸、三浦、木村

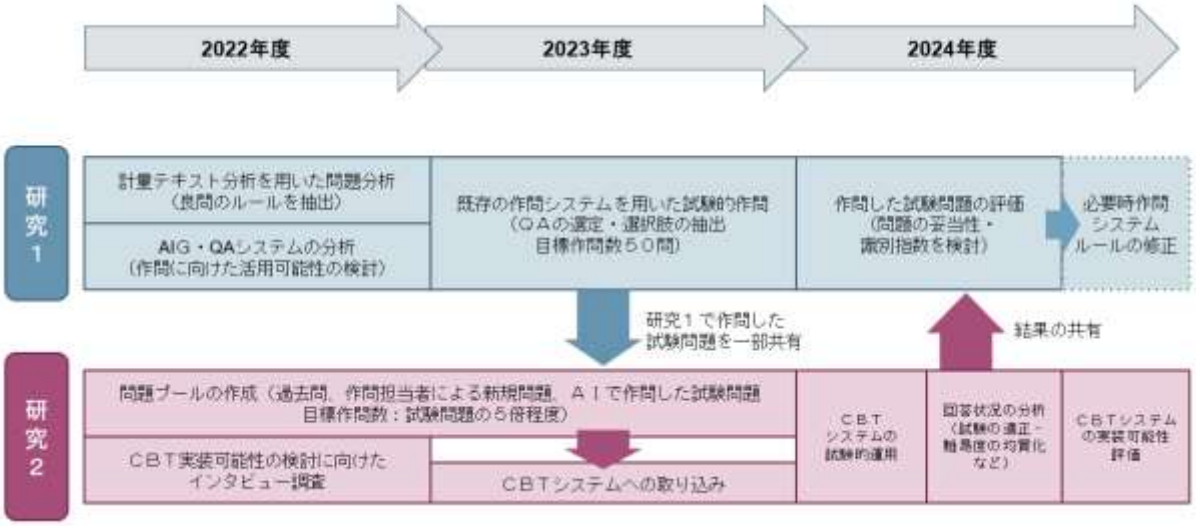
A. 研究目的

本研究は、保健師助産師看護師国家試験（以下「看護師等国家試験」）において、ICT・AI 技術等を活用した具体的な作問システムを検討すること(研究 1)、また看護師等国家試験におけるコンピュータを活用した試験（CBT：Computer-based Testing）の実装に向けて、CBT システムの試用版を開発・試行し、試験問題の妥当性について難易度、識別指数、IRT スコア等から評価すると共に、出題形式、問題管理システム、受験者側の受容性に関する調査を行い CBT 導入に向けた課題を明示すること(研究 2)を目的とした。

看護師等国家試験については、令和 3 年医道審議会報告書（看護師等国家試験制度改善検討部会）において、災害や感染症等の危機管理の観点からコンピュータ活用の必要性が示され、本研究の目的は正に喫緊の課題である。特に、膨大な問題を作問する人的負荷の軽減と問題の質の担保は、国家試験において重要な課題である。そこで、研究 1 では近年多様な分野での応用が著しい ICT・AI 技術の看護師等国家試験問題作問への活用を検討し、具体的に作問を試みた。さらに、研究 2 では将来の CBT 導入を視野に、CBT システムの実装可能性と課題を探り、基盤となるデータの創出を試みた。なお、看護師国家試験は必修問題、一般問題、状況設定問題の 3 種類で構成されるが、本研究は、基礎段階にあることから、タクソノミー I 型（単純想起問題）の必修問題に焦点をあてた。

本研究は、図 1 に示す通り令和 4（2022）年度から 6（2024）年度までの 3 年計画で進行した。

図１　本研究の３年間の全体計画



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．計量テキスト分析を用いた問題分析

1)分析対象

過去 10 年分の看護師国家試験必修問題（問題文と選択肢）

2)方法

先行研究（令和元年度厚労科研林班）で示された基準に則り、正答率、識別指数等から過去問題を「良問」と「要改善問題」に分類し、問題文と選択肢をデータとして計量テキスト分析を行い、共起ネットワーク、抽出語を比較し、「良問」に特徴的なルールを試みた。分析には KH Coder 等を用いた。

2．良問と要改善問題の評価基準の作成と良問ルールの抽出、ルールブックの開発

1)分析対象

1.の 1）に同じ（国試必修問題の過去問 10 年分）

2)方法

全 500 問の正答率、識別指数を元に良問と要改善問題の評価基準を再設定した。新たな基準により分類された結果を元に、出題基準の小項目ごとに良問と要改善問題の比較を行い、良問の特徴（良問ルール）を抽出することで、作問の手引となるルールブックの作成を開始した。なお、各問題の大項目・中項目・小項目の区分については、国家試験に関わる書籍情報に基づき、研究者らで振り分けた。

3．既存の大規模事前学習済み言語モデルの活用可能性の検討

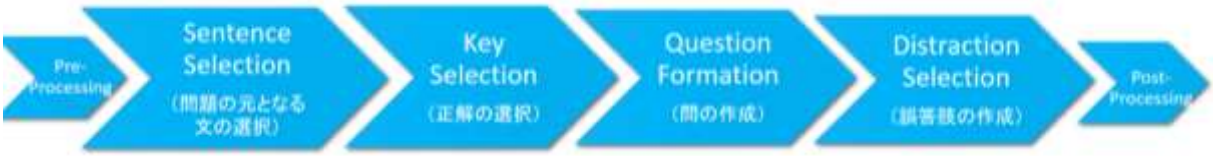
多肢選択肢問題の作成には図 2 に示すプロセスがあり、QA（Question Answering)システム [1)]は、Sentence Selection、Key

Selection、Question Formation に該当し、AIG（Automatic Item Generation）[2)3)4)]は、より広範に AI を活用したシステムと考えられる。

そこで、QA、AIG を含め、既存の大規模事前学習言語モデルの適用可能性を検討した。

図 2：問題作成システムのフロー（例）

*文献 5）図 3 より改編引用：多肢選択問題(MCQ)自動生成システムの一般的なワークフロー



4. 既存の作問システムを用いた試験的作問

QA システム、AIG の他、既存の作問システムを検討し、本課題に適したシステムを選定し作問を試みた。作問システムとして、2022 年 12 月に一般公開された、ChatGPT を使用し、様々なパターンのプロンプト（指示文）を設定して作問を試みた。

5. 過去問題の分析と作問提案

1)過去問題の内容分析

1. 計量テキスト分析を用いた問題分析の研究結果をもとに、過去 10 年分の看護師国家試験問題の必修問題 500 問について、正答率、識別指数から設定した問題の評価基準（良問：評価◯、易しすぎた問題：評価①、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△）と、問いの種類、選択肢の種類に基づき小項目を基準に問題分析を行った。

問題分析は研究者 1 名の初回分析後、研究者 3 名が分析内容の妥当性を評価し、全員の合意に至るまで、修正・評価を繰り返し、問題の改善提案を作成した。次に、看護師国家試験委員経験者、あるいは看護教育に従事する研究分担者、研究協力者全員で改善提案を検討し、作問の提案書となるルールブックに掲載する提案内容を検討した。検討の結果、改善提案について可能な限り普遍性ある内容にするべくさらなる内容分析を行い、作問システムのプロンプトとして活用可能なルールを抽出することとした。

2)作問ルールブックの作成

1）の結果に基づき、263 の改善提案について、さらなる内容分析を行なった。方法は改善提案の作成を担当した研究者 1 名が初回内容分析を行い、研究代表者を含む研究者 3 名が分析内容の妥当性を評価し、修正と評価を繰り返すことで合意形成を行った。その後、内容分析結果について、分析に携わらなかった研究分担者、研究協力者が分析内容の妥当性を評価し、指摘箇所の最終合意を得てルールブックの完成とした。

6. 既存の大規模事前学習済み言語モデル(ChatGPT)の活用可能性の検討

ChatGPT (gpt3.5-turbo-0301)の潜在的作問能力を評価するために、(１)の過去問題の分析によって作問が困難と判定された 4 問を選

び、その問題のトピック(洗髪、嘔吐、標準予防策、細胞内液)について ChatGPT を使って問題を生成した。ChatGPT へのプロンプトに含める情報として、多肢選択問題の構成要素である(1) 問題文 (Stem)、(2) 正解 (Key)に加え、(3) 看護学の標準的な教科書からトピックに関連する箇所の段落を抽出したテキスト(Text)の 3 種類の情報を考え、この組合せで 8 種類 (=2^3)のプロンプトを作成し、合計で 32 問 (=4 トピック×8 プロンプト)を作成した。なお、多肢選択問題のもうひとつの構成要素である誤答選択肢 (誤答肢) は、自動作問システムの先行研究における重要な研究課題であり、作問者にとっても作問過程において最も労力のかかる課題であることから、誤答肢は事前情報としてプロンプトに含めなかった。また、これら 3 種類の情報をすべてプロンプトに含めない場合は、トピックのみをプロンプト中に与えた。

プロンプトはあらかじめ与えなかった多肢選択問題の構成要素を段階的に尋ねる単純なものを使い、いわゆるプロンプト・エンジリアリングには注力しなかった。これは実験当時の ChatGPT が API を通じてブラックボックス的にしか利用できなかったため、プロンプト・エンジニアリングでチューニングしてもシステムが改修された時の動作の再現性が保証されないためである。したがって、今回は前述のようにプロンプト中に含める情報内容を制御することにより性能の変化を分析することとした。

作問した 32 問は 5.過去問題の分析と作問提案の研究チーム (専門家チーム) にフィードバックし、専門家の判断で ChatGPT が作問した問題がそのままで試験問題として使えるかどうかを判定することによって評価した。

7. 言語モデルによる誤答選択肢の生成と自動評価指標の作成

問題文と正解をプロンプトに含め、誤答肢を ChatGPT に生成させた問題の評価が一番高かったこと、また、多肢選択問題の作成において誤答肢を作る過程が一番労力を要するという作問経験者からの指摘を受け、誤答肢の生成に焦点を当てて研究を進めることとした。

3 つの大規模言語モデルを用い、モデルの訓練条件を変えて、誤答肢の生成性能の変化を調査した。性能を分析するためには評価が必要となるが、人手評価は時間とコストがかかるために、まず、既存の自動評価尺度を参考に、生成された誤答肢の良さを自動評価する尺度を考案した。自動評価はモデルに過去問題の問題文と正解を与えた時にモデルがその過去問題の誤答肢をどれくらい正確に再現できるかを計算することによっておこなう。

最初に情報検索の分野で利用されている再現率 (Recall)と精度 (Precision)を使うことを考えた。再現率は過去問題の誤答肢をモデルが生成できた割合を表わし、精度はモデルが生成した誤答肢が過去問題の誤答肢と一致する割合を表わす。過去問題 qi の誤答肢の集合を Gi、qi の問題文と正解を与えられてモデルが生成した誤答肢の集合を Si とすると qi に対する再現率(Ri)と精度(Pi)は以下の式で定義される。

$$\mathrm{R_i} = \frac{|S_i \cap G_i|}{|G_i|}, \quad \mathrm{P_i} = \frac{|S_i \cap G_i|}{|S_i|}$$

再現率、精度はモデルの出力と過去問題の誤答肢との表層的な一致に基づいて計算される。選択肢が名詞句のような短い表現であれば一致する可能性が高いが、選択肢が文のような長い表現の場合、たとえ意味的には同じであっても表層的な一致の可能性は著しく低

くなる。つまり、再現率、精度ではモデルの出力を過少評価してしまう可能性がでてくる。この問題を改善するために、我々は一致・不一致の二値判定に基づく再現率、精度を拡張し、表現の類似度に基づく再現率、精度を提案した。 まず、最近の自然言語処理の分野で標準的に用いられている埋め込み技術を使って、比較する誤答肢をそれぞれ多次元ベクトルに変換し、それらのベクトルが張る角度によって 2 つの誤答肢の意味的な類似度を計算する。モデルが生成した誤答肢と過去問題の誤答肢のすべての組合せについて類似度を計算し、最大の値をとる対を一致したと見なし、その値を再現率、精度の計算に利用する。2 つの誤答肢が完全に一致する場合は類似度が 1 となるように類似度を正規化しておけば、完全一致の場合は通常の再現率、精度を計算するのと同じになる。再現率、精度の定義における分子、つまり、モデルの出力と過去問題の誤答肢の積集合は以下のように書き直せる。

$$|S_i \cap G_i| = \frac{\sum_{j=1}^{|G_i|} \mathbb{1}(g_j \in S_i)}{|G_i|} = \frac{\sum_{k=1}^{|S_i|} \mathbb{1}(s_k \in G_i)}{|S_i|}$$

この式を使うと qi に対する類似度に基づく再現率(Rsi)、精度(Psi)は通常の再現率、精度の拡張として以下のように定義できる。ただし sim(•, •)は 2 つのベクトルの類似度を返す関数である。

$$\mathrm{R}_{\mathrm{s}i} = \frac{\sum_{j=1}^{|G_i|} \mathrm{sim}(\arg\max_{s \in S_i} \mathrm{sim}(s, g_j), g_j)}{|G_i|}$$

$$\mathrm{P}_{\mathrm{s}i} = \frac{\sum_{k=1}^{|S_i|} \mathrm{sim}(s_k, \arg\max_{g \in G_i} \mathrm{sim}(s_k, g))}{|S_i|}$$

Rs、Ps は完全一致の条件を緩和しているとはいえ、モデル出力と過去問題の誤答肢の一対比較の最大値に基づいて類似度を計算している。しかし、これらの尺度では、両方の集合の要素の対応関係について考慮していないため、システム出力の各誤答肢がそれぞれ異なる過去問題の誤答肢に対応付けられる状況とシステム出力の誤答肢がすべて同じひとつの過去問題の誤答肢に対応付けられる場合を区別できない。この問題を改善するために Rs、Ps をさらに拡張し、各組の組合せを考慮した類似度に基づく再現率(Rcs)、精度(Pcs)を提案した。基本的な考え方は、重複を許さずに作ったモデル出力と過去問題の誤答肢の組の集合の類似度の総和が最大になるような組合せを見つけ、この類似度の総和を再現率、精度の計算における分子として使うというものである。Si と Gi を節点集合、それらの節点の組を作ったときの対応関係を弧、類似度を弧の重みと考えると、この問題はグラフ理論における重み付き完全二部グラフの最大重みマッチング問題として定式化することができる。この問題を解く効率のよいアルゴリズムが知られている。節点集合 V1、V2 からなる重み付き完全二部グラフの最大重みマッチングを返す関数を MVM(V1, V2)とすると、Rcs、Pcs は以下のように定義できる。

$$\mathrm{R}_{\mathrm{cs}i} = \frac{\mathrm{MWM}(S_i, G_i)}{G_i}, \quad \mathrm{P}_{\mathrm{cs}i} = \frac{\mathrm{MWM}(S_i, G_i)}{S_i}$$

ChatGPT によって生成した問題の不適切な点として、同じような誤答肢が含まれているという指摘があったため、生成された誤答肢同士がどれくらい似ているかを表す尺度 DV (distractor variation)も用意した。qi に対する DVi は以下の式で定義される。

$$\mathrm{DV}_i = 1 - \frac{\sum_{\{(s_j, s_k) | s_j, s_k \in S_i, j \neq k\}} \mathrm{sim}(s_j, s_k)}{|S_i|}$$

DV は値が大きいほど互いに似ていない多様な誤答肢を生成していることを意味する。

以上の準備のもと、以下のように実験を設計した。まず、評価実験に使うデータとして厚労省から提供を受けた過去 10 年分の看護師国家試験の必修問題 500 問から、選択肢が

名詞句あるいは文からなるもの 366 問を選択し、これを問題の中項目の分野分布を維持しながら訓練用、開発用、評価用として 193、87、86 問に分割した。

大規模言語モデルとしては以下の 3 つを使用した。

- ChatGPT (gpt-3.5-turbo-0613)
- GPT-4 (gpt4-0613)
- Japanese Stable LM instruct alpha 7B-v2

学習方法としては、問題と答の例をプロンプト中に複数提示する few-shot 学習と訓練データを用いた微調整（fine-tuning)をおこなった。ChatGPT と GPT4 は OpenAI の API 経由で利用し、JSLM は研究者管理の GPU サーバーで動作させた。実験時点では GPT4 では微調整の機能が未提供だったため、GPT4 は few-shot 学習のみ、また JSLM は微調整のみをおこなった。ChatGPT については few-shot 学習と微調整の効果の差を比較するため、両方の学習をそれぞれおこなった。

8. 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 技術の活用策の検討：問題作成プロセス評価

1)調査対象

機縁法にて看護師国家試験委員経験者をリクルートし、15 名からの研究協力を得た（16 名から研究参加の同意を得たが、1 名が都合により同意を取り下げた)。

大規模言語モデルを使用して作成した誤答肢案を参照して誤答肢を作成する群（以下、AI アシスト有り群）8 名と、誤答肢案無しで全ての誤答肢を自ら作成する群（以下、AI アシスト無し群）7 名に分けた。さらに選出した問題 10 問を 5 問ずつに分け、AI アシスト有り群各 4 名（1G、2G）と無し群 4 名（3G）または 3 名（4G）に割り付け、各問題につき 3 誤答肢の作成を依頼した。その際、国家試験作問委員経験年数（5 年未満と 5 年以上）により対象者を層化した。

2)大規模言語モデル（LLM）を用いた看護師国家試験問題の誤答肢作成

①問題の選出

2013 年度～2022 年度の看護師国家試験必修問題 500 問のうち、良問の占める割合の高い出題基準大項目 4 項目、要改善問題が占める割合の高い大項目 6 項目（大項目の指定については以下参照）について、2021 年度に予備校で行われた看護師国家試験模試問題から、正答率、識別指数を考慮してそれぞれ 1 問ずつ合計 10 問を選出した（表 1)。この 10 問について大規模言語モデル（LLM）を利用して誤答肢案を作成した。予備校から同様に提供を受けた 2022 年度模試問題は、誤答肢を作成する大規模言語モデルのトレーニング (fine tuning、few shot learning 等)に利用した。

<u>良問が多い大項目</u>

1. 健康に関する指標/健康の定義と理解
3. 保険医療制度の基本/看護で活用する社会保障
10. 生命活動/人体の構造と機能
12. 薬物治療に伴う反応/薬物の作用とその管理

<u>要改善問題が多い項目</u>

2. 健康と生活/健康に影響する要因
9. 主な看護活動展開の場と看護の機能/主な看護活動の場と看護の機能
13. 基本技術/看護における基本技術
14. 日常生活援助技術
15. 患者の安全・安楽を守る技術/患者の安全・安楽を守る看護技術
16. 診療に伴う看護技術

表 1　問題セットの構成

問題番号	大項目番号	内容	問題セット
問 1	1	健康の定義と理解	QS1
問 2	3	看護で活用する社会保障	
問 3	10	人体の構造と機能	
問 4	12	薬物の作用とその管理	
問 5	2	健康に影響する要因	
問 6	9	主な看護活動の場と看護の機能	QS2
問 7	13	看護における基本技術	
問 8	14	日常生活援助技術	
問 9	15	患者の安全・安楽を守る看護技術	
問 10	16	診療に伴う看護技術	

②大規模言語モデルを用いた看護師国家試験問題の誤答肢作成のプロセス

本研究では指示応答型の LLM に対して、誤答肢を生成させるプロンプトを与えて、誤答肢を生成した。以下に示す 6 種類、7 モデルで誤答肢生成し、それらを集約して作問者に提示する誤答肢案を作成した。

1. Llama2-Swallow-70b-instruct-v0.1 (Swallow-2)
2. Llama3-Swallow-70B-Instruct-v0.1 (Swallow-3)
3. Llama3-Preferred-MedSwallow-70B (MedSwallow-3)
4. GPT-3.5-turbo (0613) finetuned (GPT-3.5-FT)
5. GPT-3.5-turbo (0613) (GPT-3.5)
6. GPT-4 (0613) (GPT-4)
7. GPT-4o (240513) (GPT-4o)

独自データで微調整可能なオープンソースのモデル(1~3)と GPT-3.5 については微調整をおこない、微調整ができないモデル(6、7)については訓練データからランダムに選択した 5 つの応答例を与えて文脈内学習をおこなった。GPT-3.5-turbo については微調整モデル（GPT-3.5-FT)と文脈内学習をおこなうモデル（GPT-3.5)の 2 種類を用意した。訓練データとしては、厚労省から提供された 2013 年度～2022 年度の看護師国家試験必修問題 500 問と予備校から提供を受けた 2022 年度模試必修問題 250 問を用いた。図 3 に使用したプロンプトを示す。〈…〉の空所には問題に応じて適切な文字列が埋められる。

図 3　プロンプトの例

【微調整モデル用】

指示文：N 択問題「〈問題文〉」で、正解選択肢が「〈正解選択肢〉」のとき、誤答選択肢を 4 つ考えてください。

【文脈内学習モデル用】

指示文：N 択問題「〈問題文〉」で、正解選択肢が「〈正解選択肢〉」のとき、誤答選択肢を 4 つ考えてください。

応答文：

誤答選択肢 1:「〈誤答肢例〉」
誤答選択肢 2:「〈誤答肢例〉」
誤答選択肢 3:「〈誤答肢例〉」

--- 以上を 5 回繰り返す ---

指示文：N 択問題「〈問題文〉」で、正解選択肢が「〈正解選択肢〉」のとき、誤答選択肢を 4 つ考えてください。

作問者によって利用されることを考えると誤答肢案には多様性が望まれるので、上記の 7 つのモデルに 4 つの誤答肢を生成させ、それを以下の方法で集約した。

1. 各モデルで誤答選択肢を 4 つずつ生成する。
2. 複数のモデルが重複して出力した選択肢を優先して候補とする。
3. 各モデルに由来する誤答肢数がそれぞれ 2 以上になるまで、生成された誤答

肢から候補に追加する。

表 2 に対象問題ごとに LLM が生成した誤答肢の種類数を示す。表の左から 3 列目は複数の LLM による重複があった数、4 列目は実際に作問者に提示する統合後の候補数を表す。

表 2　LLM によって生成された誤答肢数

問題	生成数	重複数	候補数
1	13	4	6
2	28	0	14
3	25	2	11
4	20	8	10
5	24	4	10
6	27	1	13
7	18	6	9
8	25	3	11
9	8	5	5
10	27	1	13
合計	215	34	102

3) ICT・AI 技術を活用した看護師国家試験作問委員経験者による誤答肢作成と評価（図 4「問題作成プロセス評価」のシェーマ）

①看護師国家試験の誤答肢作成

全グループに共通して、各設問の問題文と正答を提示し、1Gと2GにはAIが作成した誤答肢の案を 10 程度選択肢案として提示し、これらを参考として誤答肢を 3 つ作成するよう依頼した。ただし、AI が生成した誤答肢案を、自身が提案する誤答肢に必ずしも取り込まなくてもよいこととした。3Gと4Gには誤答肢の案は提示せず、自ら思考して誤答肢 3 つを作成するよう依頼した。

②作問後アンケート調査

①の作問後に各群共通して、各問題につき以下の質問をした。

Q1：誤答肢を 3 つ作成するのにどのくらい時間がかかったか（分）

Q2：誤答肢の作成は、国家試験委員として誤答肢を作成したときと比べて負担感はどのくらいであったか（5：負担感が大きい、4：どちらかといえば負担感が大きい、3：負担感は変わらない、2：どちらかといえば負担感が小さい、1：負担感が小さい　の 5 段階リッカートで回答）

対象属性（職位、専門領域、教員経験年数、看護師国家試験委員経験年数）

さらに、AI アシスト有り群に対しては、以下の追加質問をした。

Q3：AI が作成した誤答肢の案は、誤答肢の作成にあたり参考になったか（5：参考になった、4：どちらかといえば参考になった、3：どちらとも言えない、2：どちらかといえば参考にならなかった、1：参考にならなかった　の 5 段階リッカートで回答）

Q4：AI が作成した誤答肢の案は、ブレーンストーミングに役立ったか（5：役立った、4：どちらかといえば役立った、3：どちらとも言えない、2：どちらかといえば役立たかなかった、1：役立たなかった　の 5 段階リッカートで回答）

Q5：AI が作成した誤答肢の案は、対象者自身の自由な発想を阻害したか（5：阻害された、4：どちらかといえば阻害された、3：どちらとも言えない、2：どちらかといえば阻害されなかった、1：阻害されなかった の 5 段階リッカートで回答）

Q6：AI が作成した誤答肢の案で有効だったもの、採用したものはどれか（該当するものをすべて選択して回答）

なお、調査は全て Google フォームを用いて行った。

③分析方法

LLM によって生成した誤答肢案を作問者に

提示することが、(1)有効か否か、(2)作問の効率化に寄与するか、の 2 つの観点から分析を行った。

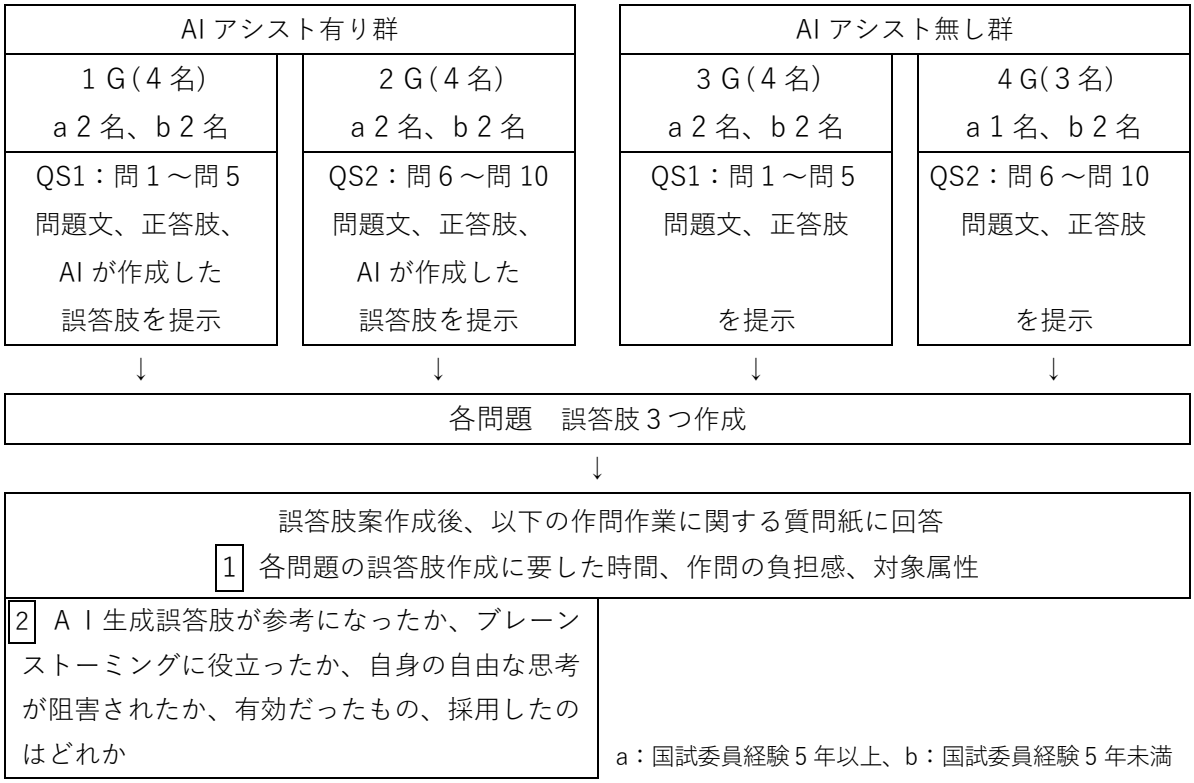
有効性については、LLM が生成した誤答肢案が作問者によって採用された採用率を主な分析の尺度として用い、(a)LLM の違い、(b)問題の違い、(c)作問者の属性の相違による採用率の違いも分析した。

また、効率化については、作問者が作問に要する時間を主な尺度として用い、作問時間が(a)問題の違い、(b)作問者の属性の違いによってどのような影響を受けるかを分析した。

4)倫理的配慮

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：24 - A042)。

図 4　「問題作成プロセス評価」のシェーマ



9. 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および CBT(Computer Based Testing)による実施の課題に関する調査

1)調査対象

学生調査の対象は日本国内の看護師養成施設に所属する、大学 3 年生以上または専門学校等（3 年課程）の 2 年生以上を対象とした。厚生労働省医政局看護課から日本看護学校協議会へ協力を依頼し、研究協力の承諾が得られた看護専門学校 5 校と、研究分担者・研究協力者の機縁による看護系大学 3 校 5 学部の学生および教員に対し参加者を募集した。学生は 399 名から応募があり、教員は 46 名から協力を得られた。

学生は試験形式の違い(筆記、CBT の 2 種

類)、および誤答肢の作成方法の違い(AI アシスト有り・無しの 2 種類)による組み合わせの 4 群に、置換ブロック法を用いて学校ごとに無作為に割り付けた。

教員は所属施設ごとにリクルートの段階で 2 つのグループに分け、1 グループ 23 名とした。各グループ 8 問ずつ、合計 16 問について、AI アシスト有り、AI アシスト無しの各誤答肢セットの評価を依頼した。

2）学生対象調査

①問題の選出と誤答肢の確定

問 1〜10：「問題作成プロセス評価」における 10 問

「問題作成プロセス評価」で対象とした問題 10 問について、AI アシスト有り群、アシスト無し群に割り付けられた対象がそれぞれ作成した誤答肢案から、「基本概念が揃っている」、「二律背反ではない」、「存在しない語句ではない」、「必ず誤りである」ことを判断基準に研究者が誤答肢を抽出し、それぞれの誤答肢セットを作成した。

問 11〜30：「問題作成プロセス評価」と同様に良問が多い大項目（1,3,10,12）と要改善問題が多い大項目（2,9,13,14,15,16）から各 2 問選択した 20 問

■AI アシスト有り誤答肢

LLM を用いて全 20 問の誤答肢案を作成し、研究者がこの誤答肢案をもとに問題ごとの誤答肢セットを作成した。

■AI アシスト無し誤答肢

予備校模試問題の誤答肢をそのまま採用した。

問 31〜50：共通問題 20 問

予備校模試問題から、大項目を限定せず正答率、識別指数を基準に良問を抽出し、AI アシスト有無の比較に使用した問題（問 1〜30）を除き、さらに類似問題は重複選択せぬよう考慮して 20 問を選出し、模試問題と同じ誤答肢を採用した。

②試験の実施方法

学生調査は対象施設（看護学校あるいは大学）ごとに、設定した日時に行った。筆記試験、CBT それぞれ別教室で実施し、解答時間は 50 問を 60 分以内とした。筆記試験は、看護師国家試験の問題冊子と同じ形式で作成した問題冊子を配布し、マークシートに解答を記入するよう依頼した。CBT はプラットフォームとして TAO を使用し、学生が所属する対象施設が保有する端末または学生が保有する端末（ノート型 PC、あるいはタブレット）からインターネットを経由して TAO システムにアクセスし解答する形式とした。CBT での解答は、筆記試験と同様に、任意の問題から解答可能とした。問題一覧から各問題に自由に遷移することができ、気になる問題等をチェックするブックマーク機能を使用可能とした。また、解答の記録やメモのために A4 サイズのメモ用紙を 1 枚配布した。問題への解答終了後に全員に自記式質問紙により、学年等の基本属性、学校内外での CBT 受験経験、パソコン等の情報機器の所持・使用状況、解答形式に対する評価についての回答を依頼した。質問紙への回答終了後に、試験の解答と解説を配布した。

③サンプルサイズ

看護師国家試験の必修問題の平均点は 45 点(50 点満点)程度であり、そのような平均点になる 1 問あたりの正答率は 90%である。共通問題の正答率、試験の媒体(紙または CBT)を考慮したうえで、作問時の LLM による支援の有無により、正答率が 10%低く(正答率 0.8、オッズ比 0.44)なった場合に検出力 0.8、

α エラー0.05 で検出するためにロジスティック回帰分析を行うとすると、必要サンプルサイズは 394 となる。
本研究では試験媒体、作問時の LLM による支援の有無の組み合わせで 4 群を割り当てるため、必要サンプルサイズを満たしかつ、きりの良い 400 名(各群 100 名)を対象数として設定した。

④分析方法

得られた解答から、作問時の AI アシストの有無別、試験形式別に、正答率、識別指数を算出した。作問時の AI アシストの有無による正答率の比較は、AI アシストの状況がことなる 30 問（問 1 から問 30）の合計および個別の問題について、試験形式、共通問題の正答数を共変量としたロジスティック回帰分析により行った。個別の問題の正答率の比較は検定の多重性を Holm 法にて調整した。

また、共通問題の得点、解答形式に対する評価、試験後の疲労感を解答形式間で比較し、Mann-Whitney の U 検定を行った。

3)教員対象調査

①問題の選出

学生対象調査で使用した、AI アシスト有り・無しの比較用問題（問 1～30）のうち、AI アシスト有り・無しで誤答肢が 2 つ以上一致または類似した問題を除外した 16 問を選出した。

②AI アシスト有り・無しの誤答肢の評価

対象者 1 名につき指定の問題 8 問について、AI アシスト有り・無しの各誤答肢セットを比較し、魅惑肢としてより適切な誤答肢セット（「どちらも適切」「どちらも不適切」の評価も可とした）とその理由について回答を依頼した。なお、対象者には提示した 2 つの誤答肢セットのうち、どちらが AI アシスト有り、または無しで作成されたものかは示さなかった。誤答肢セットの評価の後、対象属性（学校種別、専門領域、職位、教員経験年数、看護師国家試験委員経験年数）を問う設問への回答を依頼した。

以上の調査は、全て Google フォームを用いて行った。

③分析方法

得られた回答から、AI アシスト有りの誤答肢セットと AI アシスト無しの誤答肢セットの評価結果を集計し、いずれの誤答肢セットの評価が高いか、その評価の理由も含めて分析を行った。また、評価結果と出題基準における大項目との関係についても検討し、AI アシスト有りの誤答肢セットの評価と大項目との間に何らかの傾向がみられるかを検討した。

4)倫理的配慮

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：24-AC072)。

C．研究結果

1．過去問題の分析

1)計量テキスト分析を用いた問題分析

まず看護師国家試験過去 10 年分の必修問題 500 問のデータ（問題文、解答選択肢、正答、出題基準における該当する大項目、正答率、識別指数）を入手し、正答率・識別指数から、先行研究（令和元年度厚労科研林班報告書[6]）の評価基準に則って良問等の評価を行った（表 3)。また、各問題の問題形式(what・how・why)・解答形式（名詞句・数値・文・図式）の分類基準を定め、500 問を基準に基づき分析した。

次に、問題文と解答選択肢のテキストデータを対象に、評価別（◯、①、②、③、△の 5 分

類）に計量テキスト分析（KH Coder を使用）を行った。

問題の評価ごとに共起ネットワークを作成し、語の出現度、語同士のつながりを図式化した（図 5～図 9）。円の大きさは語の出現頻度を示す。

分析の結果、評価ごと（5 分類）の問題数が限られること、また出題基準の項目ごとの問題数としては少数となることから、共起語の種類と出現数が限定される結果となり、良問（◯評価）とそれ以外の問題（①、②、③、△）との相違の傾向はつかめなかった。

表 3　先行研究の基準による問題評価結果

評価		正答率％	識別指数	問題数
○	良問	90～96.15	0.2 以上	81
①	易しすぎる問題	99 以上	0.1 以下	69
②	良問だが難問	90 未満	0.22 以上	122
③	あまり適切ではない問題 識別指数 0 以下を含む	90 未満	0.15 以下	19
△	非該当問題			209
	合計			500

図 5　共起ネットワーク：良問（○）81 問

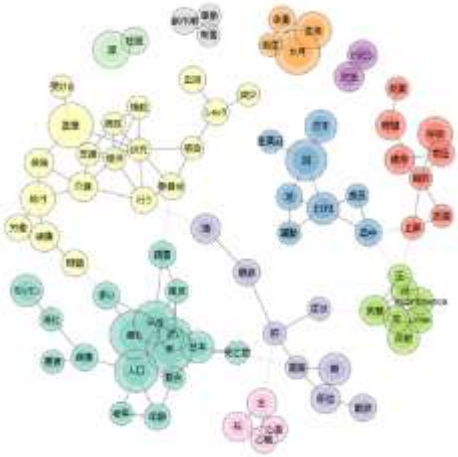


図 6　共起ネットワーク：易しすぎる問題(①)69 問

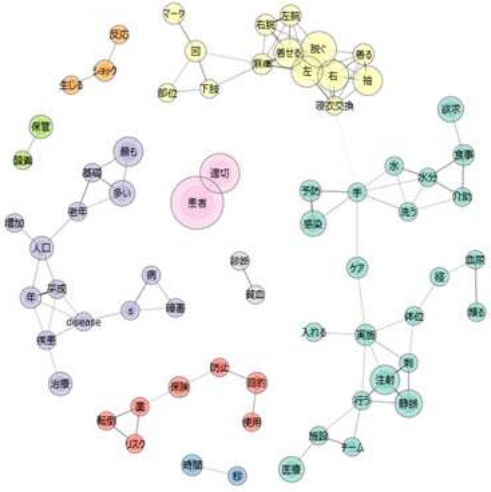


図7　共起ネットワーク：良問だが難問（②）122 問

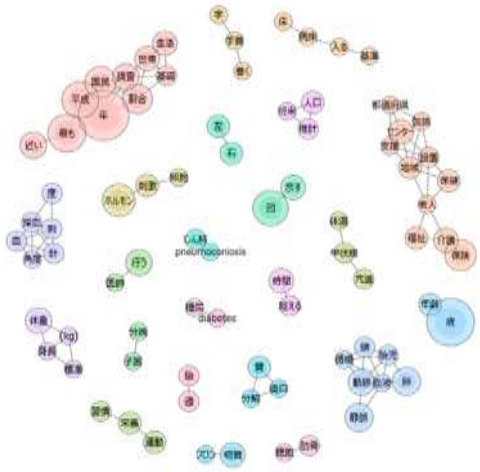


図8　共起ネットワーク：非該当問題（△）209 問

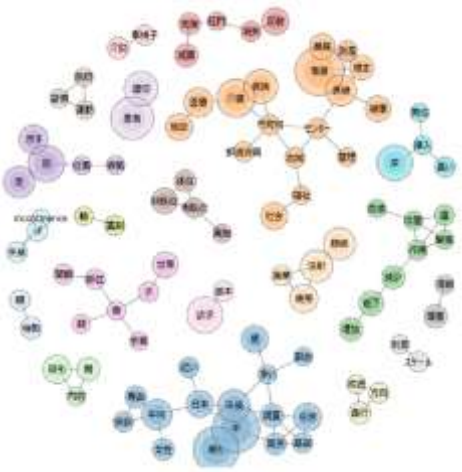
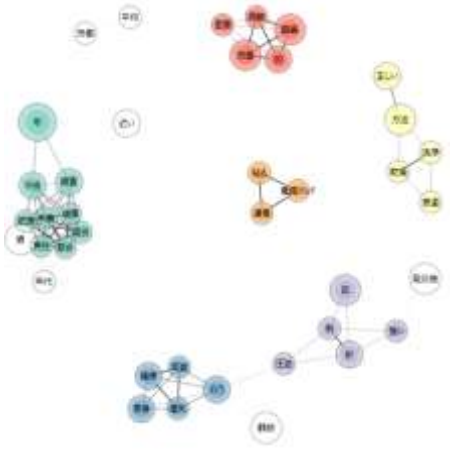


図9　共起ネットワーク：あまり適切でない問題（③）19 問



2. 問題評価基準の再検討と大項目別評価集計

全 500 問の正答率・識別指数のプロットを図 10 に示す。これより、評価②と③の間の問題（正答率 90%未満かつ識別指数 0.15 より大きく 0.22 未満）や良問より正答率が高い問題（かつ識別指数 0.2 以上）など、「△：非該当問題」の中には様々な位置にある問題があることが判明した。改善により良問となり得る問題を明確化するために、評価の基準値（区分の閾値）を再検討し、必修問題 500 問を改めて評価し可視化した（表 4、図 11）。

その結果、「良問だが難問（②）」のうち正答率が比較的高い（90%に近い）問題で、かつ識別指数の高い問題(0.3 程度以上等)について、易化を図ることで、正答率 90%以上（かつ識別指数 0.2 以上）となるように、また「非該当問題（△）」のうち正答率が高く（95~97%程度等）識別指数が比較的高い（0.2 に近い）問題について、難化を図ることで識別指数 0.2 以上（かつ正答率 90%以上）となるように、問題や解答選択肢を工夫することで良問が増えると考えられた。

次に、出題評価基準の大項目別に評価数を集計した結果（表 5)、良問の比率が高い（30%以上）大項目は、1.健康に関する指標・健康の定義と理解、3.保険医療制度の基本・看護で活用する社会保障、10.生命活動・人体の構造と機能、12.薬物治療に伴う反応・薬物の作用とその管理であった。

図 10　過去問題 500 問の正答率・識別指数プロット（○、①～③ の詳細は表 3 参照）

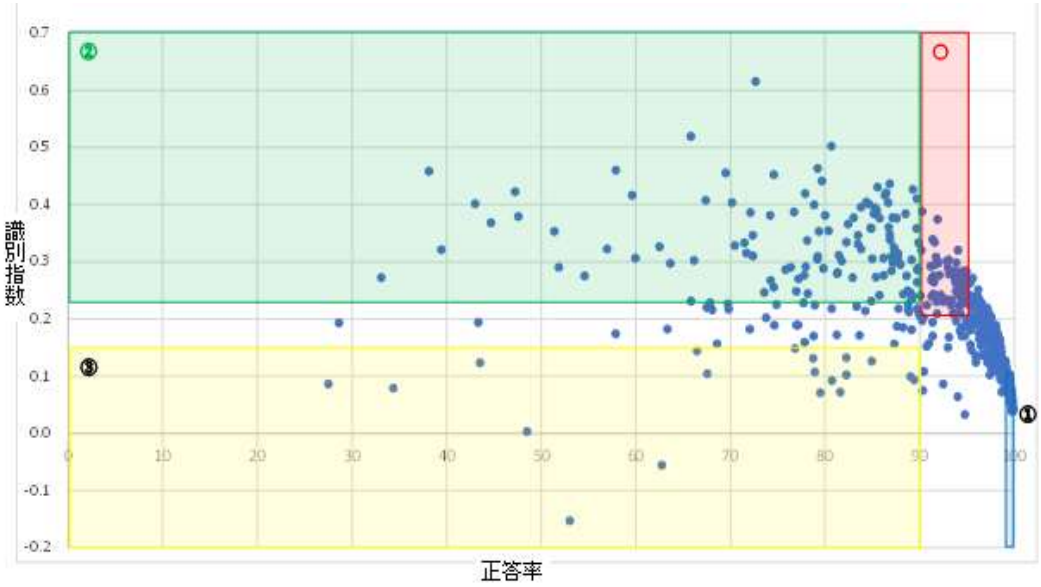


表 4　新たな基準による問題評価結果

評価		正答率%	識別指数	問題数
○	良問	90～99 未満	0.2 以上	100
①	易しすぎる問題	99 以上	−	89
②	良問だが難問	90 未満	0.2 以上	132
③	あまり適切ではない問題 識別指数 0 以下を含む	90 未満	0.2 未満	36
△	非該当問題	90～99 未満	0.2 未満	143
	合計			500

図 11　過去問題 500 問の正答率・識別指数プロット　（○、①～③、△の詳細は表 4 参照）

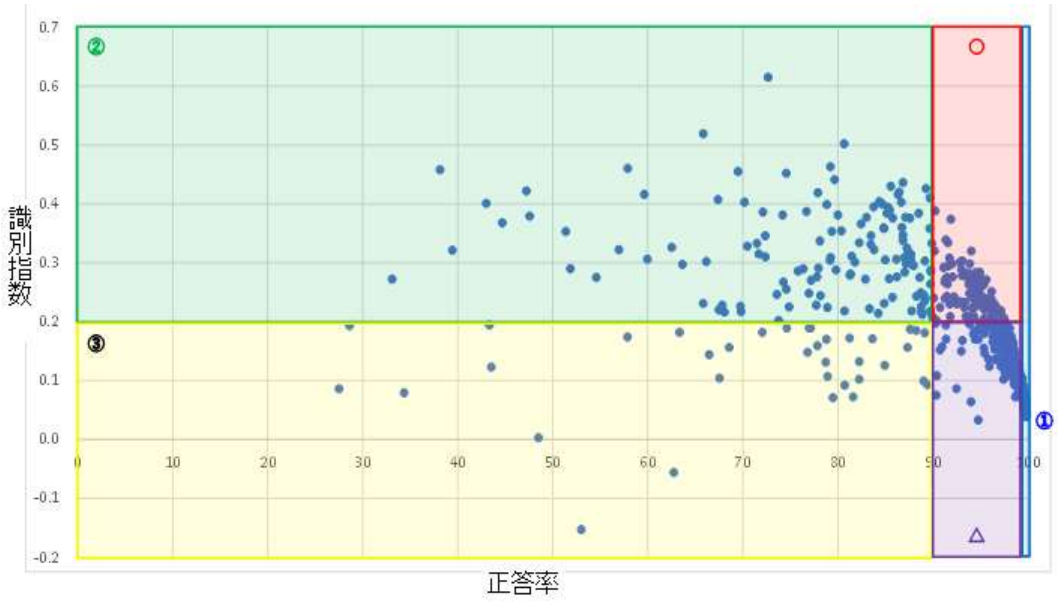


表 5　大項目別の各評価数。下段青字は項目内の比率

各大項目名の数字（H22,H26,H30）は出題基準の改定年を示す

大項目＼評価	○	①	②	③	△	計
1. 健康に関する指標 H22/H26 健康の定義と理解 H30	13 34.2%	4 10.5%	9 23.7%	2 5.3%	10 26.3%	38
2. 健康と生活 H22/H26 健康に影響する要因 H30	5 18.5%	2 7.4%	12 44.4%	3 11.1%	5 18.5%	27
3. 保険医療制度の基本 H22/H26 看護で活用する社会保障 H30	6 31.6%	1 5.3%	5 26.3%	-	7 36.8%	19
4. 看護の倫理 H22/H26 看護における倫理 H30	2 18.2%	2 18.2%	3 27.3%	1 9.1%	3 27.3%	11
5. 関係法規 H22/H26 看護に関わる基本的法律 H30	-	1 9.1%	1 9.1%	1 9.1%	8 72.7%	11
6. 人間の特性 H22/H26/H30	1 10.0%	2 20.0%	2 20.0%	1 10.0%	4 40.0%	10
7. 人間の成長と発達 H22/H26 人間のライフサイクル各期の特徴と生活 H30	11 22.9%	6 12.5%	19 39.6%	2 4.2%	10 20.8%	48
8. 患者と家族 H22/H26 看護の対象としての患者と家族 H30	-	1 33.3%	1 33.3%	-	1 33.3%	3
9. 主な看護活動展開の場と看護の機能 H22 主な看護活動の場と看護の機能 H26/H30	2 8.3%	3 12.5%	7 29.2%	2 8.3%	10 41.7%	24
10. 生命活動 H22/ H26 人体の構造と機能 H30	17 32.1%	5 9.4%	17 32.1%	5 9.4%	9 17.0%	53
11. 病態と看護 H22/ H26 主要疾患と看護 H22(大項目 12) 疾患と徴候 H30	15 19.7%	12 15.8%	30 30.3%	7 9.2%	19 25.0%	76
12. 薬物治療に伴う反応 H22(大項目 13) 薬物治療に伴う反応 H26 薬物の作用とその管理 H30	11 36.7%	2 6.7%	8 26.7%	1 3.3%	8 26.7%	30
13. 基本技術 H22(大項目 14) 基本技術 H26 看護における基本技術 H30	2 9.5%	6 28.6%	5 23.8%	-	8 38.1%	21
14. 日常生活援助技術 H22(大項目 15) 日常生活援助技術 H26/H30	6 16.2%	11 29.7%	2 5.4%	1 2.7%	17 45.9%	37
15. 患者の安全・安楽を守る技術 H22(大項目 16) 患者の安全・安楽を守る看護技術 H26/H30	1 4.0%	12 48.0%	4 16.0%	1 4.0%	7 28.0%	25
16. 診療に伴う看護技術 H22(大項目 17) 診療に伴う看護技術 H26/H30	8 11.9%	19 28.4%	14 20.9%	9 13.4%	17 25.4%	67
計	100 20.0%	89 17.8%	132 26.4%	36 7.2%	143 28.6%	500

また、各問題の問題形式（what：「何」を尋ねる、how：方法を尋ねる、why：根拠を尋ねる）・解答選択肢形式（名詞句、数値、文、図・写真）を調べ、形式別の評価数を集計した（表6、表7）。その結果、問題形式としてはwhatを尋ねる問題では「良問だが難問（②）」、「非該当問題（△）」が比較多く、how/whyを尋ねる問題では「易しすぎる問題（①）」が多かった。解答選択肢形式としては、文を選択肢としている問題では「易しすぎる問題（①）」が多く、数値を問う問題では「良問だが難問（②）」が比較的多かった。

表6　問題形式別評価数　（下段は形式内の比率）

問題形式＼評価	○	①	②	③	△	計
what	87 22.8%	52 13.6%	109 28.5%	25 6.5%	109 28.5%	382
how/why	13 11.0%	37 31.4%	23 19.5%	11 9.3%	34 28.8%	118
計	100 20.0%	89 17.8%	132 26.4%	36 7.2%	143 28.6%	500

表7　解答選択肢形式別の評価数　（下段は形式内の比率）

問題形式＼評価	○	①	②	③	△	計
名詞句	71 22.5%	59 18.7%	75 23.7%	19 6.0%	92 29.1%	316
数値	22 21.6%	6 5.9%	36 35.3%	11 10.8%	27 26.5%	102
文	4 6.6%	22 36.1%	15 24.6%	4 6.6%	16 26.2%	61
図・写真	3 14.3%	2 9.5%	6 28.6%	2 9.5%	8 38.1%	21
計	100 20.0%	89 17.8%	132 26.4%	36 7.2%	143 28.6%	500

3. 良問ルールの抽出と作問ルール提案

国家試験出題基準の小項目ごとに類似問題をまとめ、設問ごとに評価と問題形式、解答選択肢形式の組み合わせを分析した。その分析結果をもとに、良問ルールの抽出と作問ルールの提案書（作問ルールブック）の作成を試みた。

分析フォームを表 8 に示す。

表 8　過去問題の改善案と作問提案の表形式

出題基準		過去問題データ					選択肢のタイプ	問いのタイプ	評価	コメント	改善案	作問例
中項目	小項目	出題年	問題番号	問題文	選択技(1〜4)(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		

全 500 問について、表 8 に示す視点で分析を行った。各設問について、研究分担者と協力者が分析結果を記載した後（一次分析）、研究代表者がすべての問題を再分析し（二次分析）、その後研究代表者、分担者、協力者による複数回のディスカッションを経て改善案（ルール）抽出の合意形成を行った。2022 年度末時点で 500 問のうち 450 問の一次分析を終え、200 問の二次分析、100 問の合意形成に至った。結果の一例を表 9 に示す。

表 9　分析結果の例：大項目 1．健康に関する指標／健康の定義と理解

出題年	問題番号	問題文	選択肢1	選択肢2	選択肢3	選択肢4	正答	選択肢のタイプ	問いのタイプ	評価	コメント
2015	A001	日本の将来推計人口で2020年の65歳以上人口が総人口に占める割合に最も近いのはどれか。	15%	30%	45%	60%	2	数値	what	②	数値：良問だが難問、5年後の推定が難しいため数値間隔の設定が狭い可能性あり
2020	B009	平成29年(2017年)の日本の人口推計で10年前より増加しているのはどれか。	総人口	年少人口	老年人口	生産年齢人口	3	名詞句	what	①	名詞句：易問
2022	B001	平成29年(2017年)推計による日本の将来推計人口で令和47年(2065年)の将来推計人口に最も近いのはどれか。	6,800万人	8,800万人	1億800万人	1億2,800万人	2	数値	what	②	数値：間隔の設定が狭い可能性あり

【改善案】
- 「増加している'人口区分'」を問うのは簡単すぎる
- 将来人口推計に関する問いは具体的数値で問うのは難易度が高い。
- 一方で、具体的数値を問うても正答率は7割を超えている
  ⇒数値を問う問題を出し続けることで周知され◯となる可能性あり
- 数値は明確な差をつけて選択肢設定

【作問例】
Q1:◯年推計による日本の将来人口で、×年の将来推計人口に近いのはどれか。
Q2:日本の将来推計人口で◯年の老年人口（年少人口、生産年齢人口のほうが難易度上がる予測）の割合で最も近いのはどれか。

4. 作問システムの検討

1)問題の自動生成システムに関する研究動向調査

問題の自動生成に関するサーベイ論文[7)-10)]の調査に加え、最近の問題の自動生成に関する研究動向の調査をおこなった。その結果、従来から活発に研究されている QA システムのための大規模データセットを問題生成用に流用し、大規模事前学習済み言語モデルを利用した手法が主流であることがわかった。本研究課題で対象としている多肢選択問題においては、誤答肢の生成が主要な研究課題のひとつとなるが、先行研究で対象としている問題形式は穴埋め形式のものが多く、誤答肢の生成についても穴の位置になる語を言語モデルで推定し、その推定確率を元に生成する考え方が主流であった。本研究課題の分野において利用可能なデータ量で、先行研究で提案されているような手法が適用可能か、教科書データなどを入手し、引き続き検討する。

2)作問システム開発のためのテキストデータ入手

作問システム構築に使用するため、出版社から看護系の教科書 54 冊分、1 冊あたり 1800 万字程度のテキストデータを入手し、データベースの構築を開始した。今後は一般公開されている医療系テキストデータとさらなる看護系教科書、関連する医学書、薬学関係の書籍、雑誌のテキストデータを入手しデータベースを構築する。

5. ChatGPT による問題作成試行

AI を利用した大規模事前学習済み言語モデルである ChatGPT による看護師試験問題の作成を 3 つのパターンのプロンプト（A、B、C）で試みたところ、以下のような作問の結果が示された。（*イタリック*は ChatGPT の提示）

パターン A：キーワードを与え、問題文と 4 つの解答選択肢を作成

キーワード：(高齢者)（発症頻度）

問題文：高齢者に多い疾患はどれか。

①心疾患　②糖尿病　③喘息　④小児麻痺

正解選択肢：①心疾患

解説：高齢者に多い疾患は、心疾患であるとされています。高齢者の発症頻度が高い疾患としては、心疾患、糖尿病、高血圧などが挙げられます。したがって、①心疾患が正しい答えであると考えられます。

パターン B：キーワードと正解選択肢を与え、問題文と不正解の選択肢 3 つを作成

キーワード：(医師の指示)（看護師）(業務)

正解選択肢：静脈内注射

問題文：医師の指示に従って、看護師が行う業務はどれか。

①静脈内注射　②指導　③診断　④手術

正解選択肢：①静脈内注射

解説：医師の指示に従って、看護師が行う業務には、静脈内注射があります。静脈内注射は、患者さんの静脈に薬剤を注射することを指します。この業務は、医師の指示に従って看護師が行います。また、看護師は、指導や医療機器の管理など、さまざまな業務を行うことができますが、診断や手術などは医師が行う業務であり、看護師が行うことはありません。したがって、①静脈内注射が正しい答えであると考えられます。

パターン C：問題文と正解の選択肢を 1 つ与え、不正解の選択肢を 3 つ作成

問題文：患者とのコミュニケーションで適切なのはどれか。

正解選択肢：否定的感情の表出を受けとめる。

問題文：患者とのコミュニケーションで適

切なのはどれか。
①否定的感情の表出を受けとめる。
②空気を読む。
③期待感を持たせる。
④自分の意見を強調する。
正解選択肢：①否定的感情の表出を受けとめる。
解説：患者とのコミュニケーションでは、相手の感情を理解し、尊重することが重要です。患者さんが、治療や病気に対して否定的な感情を抱いている場合があります。その場合、看護師は、患者さんの否定的な感情を受け止めることが適切です。患者さんが話したいことを聞き、理解することで、患者さんが感じるストレスや苦痛を和らげることができます。したがって、①否定的感情の表出を受けとめる。が正しい答えであると考えられます。

6. 過去問題の分析と作問提案

1)過去問題の内容分析

問題分析では最終的に、全 16 の大項目、50 の中項目、269 の小項目（H30 基準）で構成される必修問題の小項目のうち、過去 10 年間で実際に出題された 200 の小項目に対して、258 の改善提案を作成した。

2)作問ルールブックの作成

1)の結果に基づき、258 の改善提案について、さらなる内容分析を行ない、200 の小項目に対して 147 の内容分析結果（ルール）が抽出された。表 10 に改善案内容分析の例を示す。

表 10 看護師国家試験過去問題　改善案内容分析（抜粋）

大項目 1. 平成 22/平成 26 健康に関する指標、平成 30 健康の定義と理解　　※は詳細非公開

中項目	小項目	出題年	問題文 選択肢　○は正答	評価※	改善提案※	作問例※	改善案内容分析※
B. 人口動態(H22)、A. 人口静態・人口動態(H26)、B. 健康に関する指標(H30)	出生、出生と死亡の動向	2013	日本の平成 22 年(2010 年)における母の年齢階級別出生率が最も高いのはどれか。 1　20～24 歳 2　25～29 歳 ③　30～34 歳 4　35～39 歳 5　40～44 歳	●	・出生数を問う設問は…。 ・母親の出産年齢について問う場合は…。 ・合計特殊出生率については…。	・日本の…における…はどれか。	・出生数については、…を…で問う。 ・合計特殊出生率については、…として出題する。
		2014	日本の平成 23 年(2011 年)における出生数に最も近いのはどれか。 1 55 万人 ②105 万人 3 155 万人 4 205 万人	●			
		2015	日本の平成 24 年（2012 年）における合計特殊出生率はどれか。 1 0.91 ②1.41 3 1.91 4 2.41	●			
		2021	平成 30 年（2018 年）の日本の出生数に最も近いのはどれか。 1 60 万人 ②90 万人 3 120 万人 4 150 万人	●			

3)既存の大規模事前学習済み言語モデル(ChatGPT)の活用可能性の検討

4 つのトピックについて 8 種類のプロンプトを用いて ChatGPT で生成した問題を（1）で過去問題を分析したメンバーにより評価した結果、問題として成立していると判断され

たものは、問題文と正解を与えてモデルが誤答肢を生成した以下の 1 問だけであった。

問題：成人の体重に占める体液の割合で最も高いのはどれか。(所与)

A. 細胞内液（正解; 所与）
B. 血液
C. リンパ液
D. 細胞外液

この結果、および誤答肢の作成が問題作成において最も負担が大きいという作問経験者からフィードバックを考慮し、問題文と正解を与えて誤答肢の候補を生成する課題に優先して取り組むこととした。

4）言語モデルによる誤答肢の生成

実験に使用したモデルについて評価用データを用いて我々の提案した尺度で評価した結果を表 11 に示す。表中の gpt3.5-few、gpt3.5-FT はそれぞれ ChatGPT に few-shot 学習、微調整を適用したモデルである。また太字になっている数字は同じ行中の最大値を示す。

表 11　誤答肢生成の評価結果

	gpt4	gpt3.5-few	gpt3.5-FT	JSLM
R	0.147	0.155	**0.178**	0.101
P	0.088	0.093	**0.107**	0.060
R_s	0.903	0.906	**0.909**	0.892
P_s	0.894	0.894	**0.896**	0.877
R_{cs}	0.901	0.903	**0.905**	0.886
P_{cs}	0.540	0.542	**0.543**	0.532
DV	0.116	0.116	0.116	**0.122**

この結果から以下の知見を得た。

- gpt3.5-few と gpt3.5-FT の結果を比較すると few-shot 学習に比べて微調整の効果が大きい。今回、微調整に使用した訓練データが 193 問と小規模であることを考えると訓練データの量を増やすことによってさらに性能の改善が期待できる。また、gpt3.5-FT は gpt4 の数値も上回っていることは微調整のさらなる可能性を示唆している。
- JSLM の性能は一貫して他のモデルより低いが、これはモデルサイズによるものが大きいと考えられる。OpenAI は ChatGPT や GPT4 のパラメタ数については公開していないが、JSLM の 70 億に比べると 2 桁以上は大きいと推測できる。自前のモデルを構築するにしても、より大きいモデルでの性能評価が必要である。
- 表 12 に評価データ中で選択肢が文であるような問題(11 問)のみの評価結果を示す。表からわかるとおり、選択肢が文の場合には、通常の再現率、精度はまったく機能していない。類似度を考慮することは意義あることだと言える。

表 12 誤答肢生成の評価結果（文選択肢）

	gpt4	gpt3.5-few	gpt3.5-FT	JSLM
R	0.000	0.000	0.000	0.000
P	0.000	0.000	0.000	0.000
R_s	**0.887**	0.884	0.881	0.874
P_s	**0.881**	0.880	0.877	0.856
R_{cs}	**0.885**	0.882	0.877	0.862
P_{cs}	**0.531**	0.529	0.526	0.517
DV	**0.115**	0.111	0.088	0.114

- DV の尺度のみ JSLM は他のモデルより数値が大きい。これは他のモデルに比べて JSLM が生成した誤答肢は互いに似ていないことを意味する。JSLM の再現率、精度が他のモデルより一貫して低いことを考えると JSLM は過去問題の誤答肢を再現することは苦手だが、より多様性のある誤答肢を生成している可能性がある。ただし、それが誤答肢として適切かどうかを判断するは専門家による評価が必要である。

7. 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 技術の活用策の検討：問題作成プロセス評価

大規模言語モデルを用いて作成した看護師国家試験問題の誤答肢案を参考に、看護師国家試験作問委員経験者による誤答肢作成と評価を依頼した。

対象者の背景を表 13 に示す。

表 13　対象者の背景　　N=15

	AI アシスト有	AI アシスト無	計
職位			
教授	7	3	10
専門学校長	1	0	1
専門学校副学長（前・現）	0	2	2
専門学校教務主任	0	1	1
病院医療安全室 副室長	0	1	1
専門分野（重複あり）			
基礎看護学	7	5	12
成人看護学	1	2	3
看護管理学	1	1	2
精神看護学	0	2	2
看護教育学	1	0	1
感染看護学	1	0	1
臨床薬理学	0	1	1
看護教員の経験年数			
1〜10 年	0	1	1
11〜15 年	0	0	0
16〜20 年	1	1	2
21〜25 年	5	3	8
26〜30 年	1	2	3
31〜35 年	1	0	1
看護師国家試験委員としての経験年数			
1〜5 年	4	4	8
6〜10 年	2	2	4
10〜15 年	2	1	3

職位は教授が最も多く、専門分野としては基礎看護領域が多かった。看護教員としての経験年数は 20 年から 25 年未満が最も多く、国家試験委員としての経験は、5 年未満が最も多く、次が 6 年から 10 年であった。

1）誤答肢案の有効性

対象とした 10 問について作問者が作成した誤答肢の異り数は 92 であった。このうち 48 は LLM が提案した誤答肢案に含まれており、採用率は 52%であった。また、作問者に提示した誤答肢案の総数は表 2 に示すとおり 102 であり、このうち 48 が採用されていたため誤答肢案の 47%が妥当と判断された。また、アンケートの Q6 の回答より、採用はしなかったが誤答肢としては妥当だと判断された 22 の誤答肢案を加えると、LLM が生成した誤答肢案の 69%（70/102）が妥当と判断されていた。

採用された 48 の誤答肢案のうち、11(22%)は、誤答肢案を参照しないグループ（AI アシスト無し群）3G、4G の作問者が作成した誤答肢と重複していた。48 の誤答肢案と各問題で実際に予備校の模試で使用された誤答肢(真の誤答肢)の重なりは、48 中 6(13%)であった。これらの真の誤答肢案も、グループ 1G、2G の作問者に採用されていた。

2）LLM の違い

表 14 に LLM 別の誤答肢案の採用率を示す。予想に反して、かなり古いモデルである Swallow-2 が最も高い採用率を示した。一方、医学テキストで学習した MedSwallow-3 は、そのベースモデルの Swallow-3 よりも採用率が低い結果となった。GPT-4 ファミリーは多くの誤答肢案を提供しているが、その採用率は GPT-3.5 ファミリーよりも低く、これも想定とは異なる結果であった。

表 14　LLM ごとの誤答肢案

LLM	誤答肢案（採用）	採用率	「真の」誤答肢
Swallow-2	23 (18)	0.78	4
Swallow-3	25 (17)	0.68	5
MedSwallow-3	22 (13)	0.59	4
GPT-3.5	23 (15)	0.65	3
GPT-3.5-FT	22 (11)	0.50	1
GPT-4	25 (11)	0.44	4
GPT-4o	26 (12)	0.46	4

誤答肢案の中には、対象とする問題の元の誤答肢（模試問題）と同じ「真の」誤答肢が 6 つあり、この 6 つの誤答肢はすべて作問者によって採用された。表 14 には各 LLM が生成した誤答肢のうち、「真の」誤答肢の数を示した。GPT-3.5 ファミリーを除くモデルは「真の」誤答肢をよく再現していた。一方、高い再現率は必ずしも高い採用率を示してはいなかった。

3)　選択肢の種類による違い

誤答肢の採用率は設問によって異なり、0.13 から 1.00 の幅があった。そこで、LLM が生成した誤答肢案が採用されやすい問題の特徴を検討した。選択肢のタイプは名詞句と文の 2 種類で、今回の問題セットの問 8 と問 10 は文の選択肢、その他は名詞句の選択肢である。これらのグループのミクロ平均採用率を計算すると、名詞句の選択肢では 0.61、文の選択肢では 0.32 となった。

QS1（問 1～問 5）と QS2（問 6～問 10）の問題セットの違いについて、平均採用率は QS1 が 0.72、QS2 が 0.38 であった。外れ値の Q1 と Q7 を取り除くと差は縮まるが、それでも 0.64（QS1）と 0.44（QS2）であった。

4）作問者の教員歴と誤答肢案採択率

作問者の採用率にも、0.13 から 1.00 まで大きなばらつきがあった。LLM が生成した誤答肢案を採用しやすい作問者の特徴を検討した結果、明らかな特徴として、経験年数の相違が挙げられた。4 人の経験の浅い群(試験委員経験 5 年未満)と 4 人の経験の長い群（同 5 年以上）についてミクロ平均採用率を計算すると、前者は 0.70、後者は 0.48 であった。経験が長い群は LLM が生成した誤答肢案の採用率が低い傾向があった。

次に、両群の問題セットの影響を検討した。経験の浅い群では、採用率は QS1 で 1.0、QS2 で 0.40 であった。一方、経験の長い群では 0.53 と 0.43 である。ここでも、経験の浅い群が質問セットの違いの影響を受けていた。

5）誤答肢案による効率化

AI アシスト有り群（G1、G2)と AI アシストなし群（G3、G4)のアンケート結果について、全問題の平均値を表 15 に、各問題の平均値を表 16 および表 17 に示す。

表 15　アンケート結果　全問題の平均値

AI アシスト	Q1	Q2	Q3	Q4	Q5
あり	7.0	2.2	3.6	3.1	2.0
なし	21.0	2.6	–	–	–

作問に要する時間(Q1)は誤答肢案を提示することにより、平均 21 分から 7 分へ、約 1/3 に短縮していた。国試試験委員として誤答肢を作成した時との作業負担の比較(Q2)についてはどちらも 3.0 を下回っていたが、AI アシスト有り群の方がその平均値が小さく、より負担を軽く感じていた（2.2 対 2.6)。また、AI アシスト有り群に尋ねた質問項目については、誤答肢案が参考になったか(Q3)、ブレーンストーミングに役立ったか(Q4)、のいずれも 3.0 を

越えており、自由な発想を阻害したか(Q5)については 3.0 を下回り、誤答肢案が作問の障害にはなっていないことが示された。

6）問題セットと所要時間

問題セット別に見ると、QS1 では AI アシスト有り群の方がわずかではあるが、より多くの時間を作問に使っていることがわかった（5.9 分と 4.8 分）。一方、QS2 については作問に要した平均時間はそれぞれ 8.0 分、37.3 分であった。時間短縮は主に QS2 によるものであった。

7）作問者の教員歴と所要時間

AI アシストの有無にかかわらず、経験が浅い群と長い群で作問平均時間を算出したところ、前者(15.9 分)は後者の(8.5 分)の約 2 倍の時間を要していた。さらに、経験の浅い群では、AI アシスト無し群(25.9 分)は AI アシスト有り群(5.9 分)の作問時間に比べ 4.4 倍かかっていたのに対し、経験が長い群では両者の差はほとんどなかった（アシスト無し群：9.1 分、アシスト有り群：8.1 分)。

作業負担(Q2)についても同様の分析を行った。経験の浅い群と経験の長い群の平均作業負担のスコアはそれぞれ 2.4 と 2.3 であった。この差は作業時間の差よりも小さい。しかし、これを作問経験別に見てみると、経験の浅い群の平均スコアは AI アシスト有り群で 2.0、AI アシストなし群で 2.9 点であるのに対し、経験の長い群ではそれぞれ 2.5、2.1 と差が少なかった。

各問題の AI 生成誤答肢と作問者による回答詳細を巻末資料表 26 に示す。

表 16　アンケート結果 各問題平均値　N=15

	AI アシスト 有り N=8	AI アシスト 無し N=7
Q1 作問時間（分）：誤答選択肢を作成するためにかかった時間		
問題1	4.0	2.8
問題2	9.5	8.3
問題3	5.0	3.5
問題4	5.5	3.5
問題5	5.5	5.8
問題6	8.3	28.3
問題7	6.0	33.3
問題8	8.5	40.0
問題9	6.0	38.3
問題 10	12.0	46.7
全問平均	**7.0**	**21.0**
Q2 作業負担感：国家試験委員として誤答肢を作成したときと比べての負担感（1：負担感小、5：負担感大）		
問題1	1.0	2.0
問題2	2.5	3.0
問題3	2.3	2.8
問題4	2.0	2.5
問題5	2.3	2.0
問題6	1.8	2.0
問題7	2.5	3.0
問題8	2.5	2.7
問題9	3.0	3.0
問題 10	2.5	2.7
全問平均	**2.2**	**2.6**

表 17　アンケート結果

AI アシスト有り群のみの質問　　N=8

	参考度 ＊1	ブレスト貢献 ＊2	阻害度 ＊3	AI 生成誤答肢 採用数／候補数 ＊4
問題1	4.8	4.5	1.8	3.0／ 6
問題2	3.8	3.8	2.3	2.8／14
問題3	4.5	4.3	1.5	2.5／11
問題4	4.5	4.3	1.3	3.3／10
問題5	4.8	4.5	1.8	3.8／10
問題6	3.8	2.8	2.8	2.3／13
問題7	1.8	1.5	2.8	0.3／ 9
問題8	2.5	1.5	2.3	1.0／11
問題9	2.3	2.3	1.5	1.8／ 5
問題 10	3.5	2.0	2.3	1.8／12
全問平均	**3.6**	**3.1**	**2.0**	

＊1 参考度：AI 作成誤答肢案は参考になったか
（1：参考にならなかった、5：参考になった）

＊2 ブレスト貢献：AI 作成誤答肢案はブレーンストーミングに役立ったか（1：役に立たなかった、5：役に立った）

＊3 阻害度：AI 成誤答肢案は自身の自由な発想を阻害したか（1：阻害されなかった、5：阻害された）

＊4 採用誤答肢数：AI 成誤答肢案のうち有効だったまたは採用した数

8. 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

1)学生対象調査

① 回答状況および参加者の特徴

割り付け結果のフローチャートを図 12 に、調査参加者の特徴を表 18 に示した。399 名の応募者のうち、15 名が当日の体調不良等で欠席であったため、参加者は 384 名であった。

調査参加者の年齢の中央値は 22 歳、女性が約 90%を占め、大学所属の者は 45%程度、学年は大学は 4 年生、専門学校は 3 年生の参加者が多かった。これらの特徴はグループ間で大きな差は認められなかった。CBT の経験については、70%程度が何らかの形で経験しており、学校では 5〜6 割、学校外の資格試験では 3〜4 割が経験していた。パソコンについては 8 割程度が自分専用または共用のものを所持しており、4 割程度は週に 1 回以上使用していた。スマートフォン・タブレット端末は、ほぼ全員が自分専用の物を所持しており、95%を超える対象者が毎日使用していた。最後に、各グループで共通の問題の得点には差はなかった。

② 作問時の LLM 支援の有無別の問題の正答率の比較

表 19 に作問時の LLM 支援の有無別の問題の正答率を比較した結果を示した。30 問合計の正答率は AI アシスト有り問題を解いたグループで 83.0%、AI アシスト無し問題を解いたグループで 84.6%であり、有意な差は認められなかった。問題ごとの正答率をみると、問 2、3、18、22、25、27 において、検定の多重性を調整しても有意な差が認められ、問 27 以外は AI アシスト有りの問題のほうが正答率が低かった。

図 12 学生対象調査の割り付けフローチャート

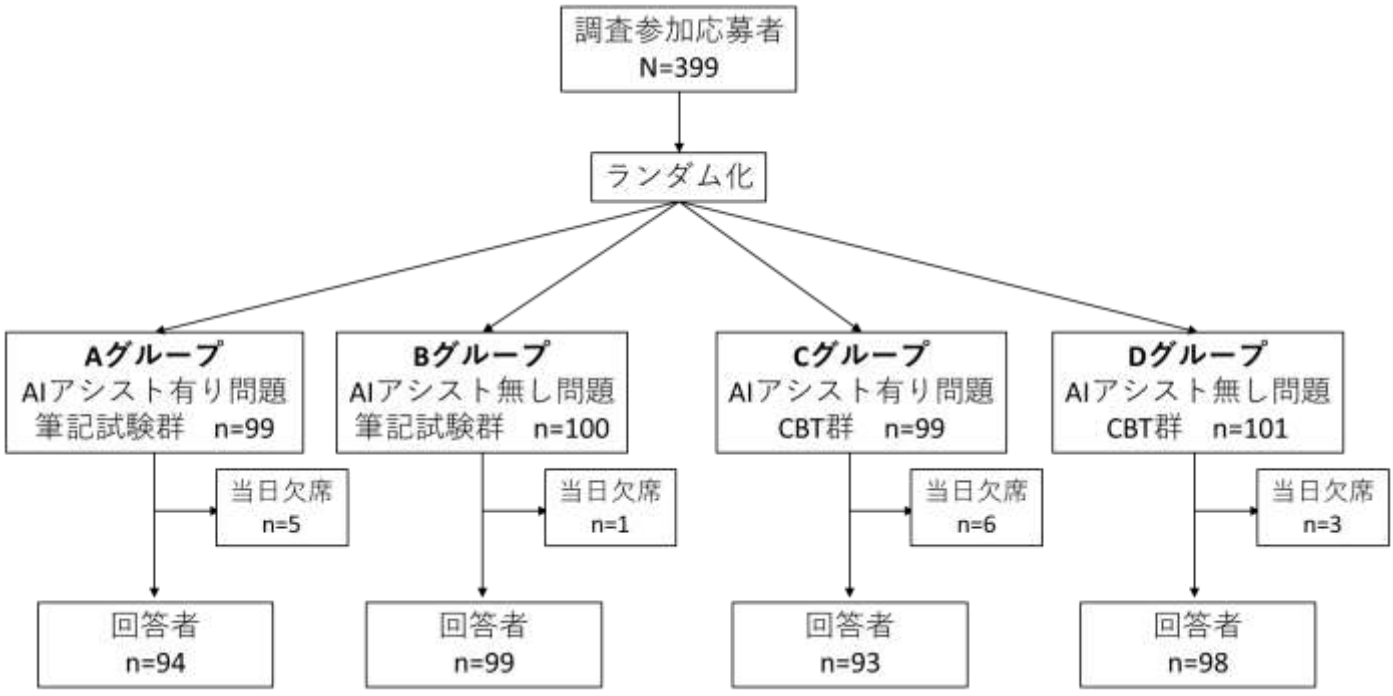


表 18 学生調査の対象者の特徴

	Aグループ n = 94	**Bグループ** n = 99	**Cグループ** n = 93	**Dグループ** n = 98
年齢	22.0 (21.0, 23.0)	22.0 (21.0, 23.0)	22.0 (21.0, 23.0)	22.0 (21.0, 23.0)
無回答	2	0	0	0
性別				
男性	12 (13%)	9 (9.1%)	6 (6.5%)	10 (10%)
女性	81 (87%)	90 (91%)	87 (94%)	88 (90%)
無回答	1	0	0	0
大学	42 (45%)	45 (45%)	43 (46%)	42 (43%)
学年				
専門学校2年生	17 (18%)	17 (17%)	15 (16%)	20 (20%)
専門学校3年生	33 (36%)	37 (37%)	35 (38%)	36 (37%)
大学3年生	5 (5.4%)	7 (7.1%)	5 (5.4%)	4 (4.1%)
大学4年生	37 (40%)	38 (38%)	38 (41%)	38 (39%)
無回答	2	0	0	0
CBTの経験	62 (67%)	64 (65%)	65 (70%)	70 (71%)
学校の試験でのCBTの経験				
あり	49 (53%)	53 (54%)	55 (59%)	57 (58%)
なし	44 (47%)	46 (46%)	38 (41%)	41 (42%)
無回答	1	0	0	0
学校外の試験でのCBTの経験				
あり	32 (34%)	35 (35%)	38 (41%)	37 (38%)
なし	61 (66%)	64 (65%)	55 (59%)	61 (62%)
Unknown	1	0	0	0
パソコンの所持				
自分専用	50 (54%)	59 (60%)	51 (55%)	50 (51%)
共用	24 (26%)	23 (23%)	23 (25%)	23 (23%)
持っていない	19 (20%)	17 (17%)	19 (20%)	25 (26%)
無回答	1	0	0	0
パソコンの使用頻度				
毎日	9 (9.7%)	13 (13%)	12 (13%)	12 (12%)
週4～6日	13 (14%)	9 (9.1%)	14 (15%)	11 (11%)
週1～3日	18 (19%)	21 (21%)	16 (17%)	15 (15%)
月1～3日	16 (17%)	17 (17%)	12 (13%)	15 (15%)
月1回未満	37 (40%)	39 (39%)	39 (42%)	45 (46%)
無回答	1	0	0	0
スマートフォン、タブレット端末の所持				
自分専用	93 (100%)	99 (100%)	92 (99%)	97 (99%)
持っていない	0 (0%)	0 (0%)	1 (1.1%)	1 (1.0%)
無回答	1	0	0	0
スマートフォン・タブレット端末の使用頻度				
毎日	89 (96%)	94 (95%)	88 (95%)	96 (98%)
週4～6日	3 (3.2%)	4 (4.0%)	2 (2.2%)	1 (1.0%)
週1～3日	1 (1.1%)	1 (1.0%)	0 (0%)	1 (1.0%)
月1～3日	0 (0%)	0 (0%)	2 (2.2%)	0 (0%)
月1回未満	0 (0%)	0 (0%)	1 (1.1%)	0 (0%)
無回答	1	0	0	0
共通問題の正答率(%)	95 (90, 100)	95 (90, 100)	95 (90, 100)	95 (90, 100)

量的変数は Median (IQR)、質的変数は n (%)を示している。

A グループ: アシストあり問題・筆記試験　　B グループ: アシストなし問題・筆記試験

C グループ: アシストあり問題・CBT　　D グループ: アシストなし問題・CBT

表 19 作問時の LLM 支援の有無別の正答率の比較

	AIアシスト有り問題		AIアシスト無し問題			95%信頼区間			
	正答率	識別指数	正答率	識別指数	オッズ比	下限	上限	P値	P値(多重比較調整)
合計	83.0%		84.6%		0.97	0.92	1.03	0.3	-
問1	98.4%	0.043	92.9%	0.214	4.13	1.20	19.28	0.039	0.777
問2	69.0%	0.511	86.3%	0.234	0.32	0.19	0.54	<0.001	0.001
問3	74.9%	0.426	85.8%	0.38	0.36	0.19	0.64	0.001	0.019
問4	81.3%	0.447	80.2%	0.52	0.92	0.51	1.66	0.793	1.000
問5	62.6%	0.468	68.5%	0.547	0.71	0.45	1.10	0.125	1.000
問6	64.2%	0.468	74.6%	0.407	0.57	0.36	0.90	0.015	0.369
問7	79.1%	0.553	76.6%	0.514	1.05	0.61	1.83	0.854	1.000
問8	63.6%	0.532	71.1%	0.594	0.62	0.38	0.99	0.047	0.884
問9	91.4%	0.149	93.4%	0.2	0.65	0.29	1.44	0.297	1.000
問10	85.6%	0.362	81.2%	0.46	1.26	0.70	2.28	0.434	1.000
問11	95.2%	0.085	93.9%	0.14	1.18	0.47	3.01	0.729	1.000
問12	96.3%	0.149	91.4%	0.254	1.93	0.73	5.52	0.198	1.000
問13	94.1%	0.128	93.9%	0.16	0.89	0.36	2.17	0.794	1.000
問14	96.8%	0.064	90.9%	0.194	2.82	1.13	8.04	0.035	0.728
問15	83.4%	0.383	78.7%	0.62	1.27	0.71	2.26	0.420	1.000
問16	81.3%	0.362	88.3%	0.26	0.50	0.27	0.90	0.023	0.524
問17	81.8%	0.468	87.3%	0.4	0.50	0.26	0.93	0.031	0.684
問18	84.5%	0.447	96.4%	0.12	0.10	0.03	0.25	<0.001	<0.001
問19	94.1%	0.149	94.4%	0.2	0.72	0.27	1.84	0.491	1.000
問20	98.9%	0.021	95.9%	0.08	3.68	0.90	24.72	0.104	1.000
問21	98.4%	0.021	99.0%	0.04	0.50	0.06	3.20	0.468	1.000
問22	64.2%	0.574	86.8%	0.28	0.20	0.11	0.35	<0.001	<0.001
問23	60.4%	0.574	59.4%	0.607	1.00	0.65	1.54	0.994	1.000
問24	97.3%	0.064	93.4%	0.1	2.53	0.93	8.02	0.085	1.000
問25	60.4%	0.383	75.6%	0.567	0.42	0.26	0.67	<0.001	0.009
問26	85.6%	0.319	86.8%	0.36	0.74	0.39	1.41	0.368	1.000
問27	79.1%	0.553	62.9%	0.621	2.42	1.47	4.05	0.001	0.016
問28	89.3%	0.34	82.2%	0.5	1.74	0.88	3.52	0.115	1.000
問29	89.3%	0.234	85.8%	0.374	1.28	0.68	2.44	0.454	1.000
問30	89.8%	0.255	82.7%	0.307	1.74	0.92	3.36	0.090	1.000

注 1：オッズ比は試験方式、共通問題の得点を共変量としたロジスティック回帰分析により推定した。1 より大きければアシストありが正答率が高いことを表す

注 2：多重比較の調整は Holm 法で行った

次に、多重性を調整しても正答率に有意な差があった問題の解答選択肢の分布を表 20 に示した。

問 2（介護保険による給付に関する問題）では、「医療機関の入院費用(12.3%)」「介護保険施設の食費(17.1%)」にも解答が分散していた。問 3(臓器の移植に関する法律における脳死の判定基準に関する問題)では、「自発呼吸」「心拍停止」は AI アシストの有無間で共通であり、解答の分散の程度も同等であった

が、AI アシスト有りの問題の「瞳孔散大」は11.8%が選択していた。問 18(ワルファリンカリウムの作用に影響をおよぼす食品に関する問題)では、AI アシスト有りの問題の選択肢「グレープフルーツ」が 15.5%選択されていた。問 22(地域包括支援センターの機能に関する問題)では、AI アシスト有り問題の「高齢者福祉サービスの提供」が 33.7%選択されており、正答率を下げていた。問 25(成人のグリセリン浣腸に関する問題)では、「浣腸液の温度は常温である」は 25.4%選択されており、正答率を下げていた。問 27(病室の環境についての問題)では、AI アシスト有りの問題の「夏季の室温は 30～32℃とする」が 1.6%と選択された割合が低かった。

表 20　正答率に差があった問題の解答分布

	AI アシスト有り問題	n = 187		AI アシスト無し問題	n = 197	
	選択肢	回答数	割合	選択肢	回答数	割合
問 2	介護保険による給付はどれか。					
正答	地域密着型サービスの費用	129	69.0%	地域密着型サービスの費用	170	86.3%
誤答肢	健康診断の費用	3	1.6%	教育訓練給付	5	2.5%
誤答肢	医療機関の入院費用	23	12.3%	訪問理髪サービス	9	4.6%
誤答肢	介護保険施設の食費	32	17.1%	家事代行サービスの費用	13	6.6%
問 3	臓器の移植に関する法律における脳死の判定基準に該当するのはどれか。					
正答	平坦脳波	140	74.9%	平坦脳波	169	85.8%
誤答肢	自発呼吸	12	6.4%	心停止	13	6.6%
誤答肢	心拍停止	13	7.0%	自発呼吸	8	4.1%
誤答肢	瞳孔散大	22	11.8%	瞳孔縮小	7	3.6%
問 18	ワルファリンカリウムの作用に影響を及ぼす食品はどれか。					
正答	納豆	158	84.5%	納豆	190	96.4%
誤答肢	牛乳	0	0.0%	牛乳	3	1.5%
誤答肢	バナナ	0	0.0%	米飯	0	0.0%
誤答肢	グレープフルーツ	29	15.5%	チーズ	4	2.0%
問 22	地域包括支援センターの機能はどれか。					
正答	介護予防ケアマネジメント	120	64.2%	介護予防ケアマネジメント	171	86.8%
誤答肢	介護認定審査	3	1.6%	金銭の管理	4	2.0%
誤答肢	医療設備の提供	1	0.5%	配食サービス	18	9.1%
誤答肢	高齢者福祉サービスの提供	63	33.7%	訪問介護の実施	4	2.0%
問 25	成人のグリセリン浣腸で正しいのはどれか。					
正答	グリセリン濃度は 50 %である。	113	57.4%	グリセリン濃度は 50 %である。	149	75.6%
誤答肢	立位で行う。	2	1.0%	立位で行う。	0	0.0%
誤答肢	浣腸液の温度は常温である。	50	25.4%	肛門から 10 cm 挿入する。	23	11.7%
誤答肢	浣腸液の酸性度は pH2-3 である。	21	10.7%	浣腸液の温度は 30 ℃とする。	24	12.2%
問 27	病室の環境で適切なのはどれか。					
正答	冬季の室温は 20～22 ℃とする。	148	79.1%	冬季の室温は 20 ～ 22 ℃とする。	124	62.9%
誤答肢	湿度は 75～80%とする	20	10.7%	夜間の騒音は 100dB 以下とする。	29	14.7%
誤答肢	夏季の室温は 30～32 ℃とする。	3	1.6%	病室の床面積は患者 1 人につき 3.0 ㎡以上とする。	24	12.2%
誤答肢	照度は 1000 ルクス以上である。	16	8.6%	昼間の病室の推奨照度は 1,000 ルクスとする。	20	10.2%

2)教員対象調査

対象者の背景を表 21 に示す。所属は看護専門学校が 71.7%、看護系大学が 28.3%であった。職位は役職なしの教員、看護教員の経験年数は 6〜10 年が最も多く 28.3%、専門分野は基礎看護学が最も多かった。看護師国家試験委員の経験者は 15.2%で、そのうち経験年数は 1〜5 年が最も多かった。学内で国家試験の指導経験が有ると回答したのは 71.7%で、その内容は、国家試験対策計画立案、ガイダンス・学習方法・文献の指導、解説（国家試験過去問、学内テスト、模試)、国家試験対策講座・勉強会・補習での指導、自らの専門分野の国家試験用講義、模試後の面談・個別指導であった。

表 21 対象者の背景　N=46

	人数	%
学校種別		
専門学校	33	71.7
大学	13	28.3
職位		
教授	1	2.2
准教授	4	8.7
専任講師	5	10.9
助教	3	6.5
助手	2	4.3
教員（役職名なし）	27	58.7
学科長	1	2.2
教務主任	1	2.2
教務部長	1	2.2
実習調整者	1	2.2
看護教員の経験年数		
1 年未満	6	13.0
1〜5 年	10	21.7
6〜10 年	13	28.3
11〜15 年	11	23.9
16〜20 年	2	4.3
21 年以上	4	8.7
専門分野		
基礎看護学	10	21.7
小児看護学	7	15.2
成人看護学	7	15.2
精神看護学	5	10.9
老年看護学	4	8.7
母性看護学	3	6.5
在宅看護論	3	6.5
看護管理学	1	2.2
地域看護学	1	2.2
助産学	1	2.2
家族看護	1	2.2
精神看護と国際、災害看護	1	2.2
看護の統合	1	2.2
看護基盤学	1	2.2
国家試験委員の経験		
有り	7	15.2
看護師国家試験委員としての経験年数		
1 年未満	0	0.0
1〜5 年	5	71.4
6〜10 年	1	14.3
不明	1	14.3
無し	39	84.8
学内での国家試験の指導経験		
有り	33	71.7
指導経験の内容（記述回答）		
・国家試験対策計画立案		
・ガイダンス・学習方法・文献の指導		
・解説（国家試験過去問、学内テスト、		
・国家試験対策講座・勉強会・補習での		
・自らの専門分野の国家試験用講義		
・模試後の面談・個別指導		
無し	13	28.3

各問題の誤答肢セットの評価結果を図 13 に示す（評価の詳細は巻末資料表 27 参照)。「AI アシスト有りの誤答肢セットの方が適切」の比率が多い問題の順に示した。その結果、「AI アシスト有り」の誤答肢セットの方が、「AI アシスト無し」の誤答肢セットより適切と判断した人の割合が高かった問題(図 13 の赤破線まで)は 16 問中有 9 問(56.3%)、「AI アシスト有り」

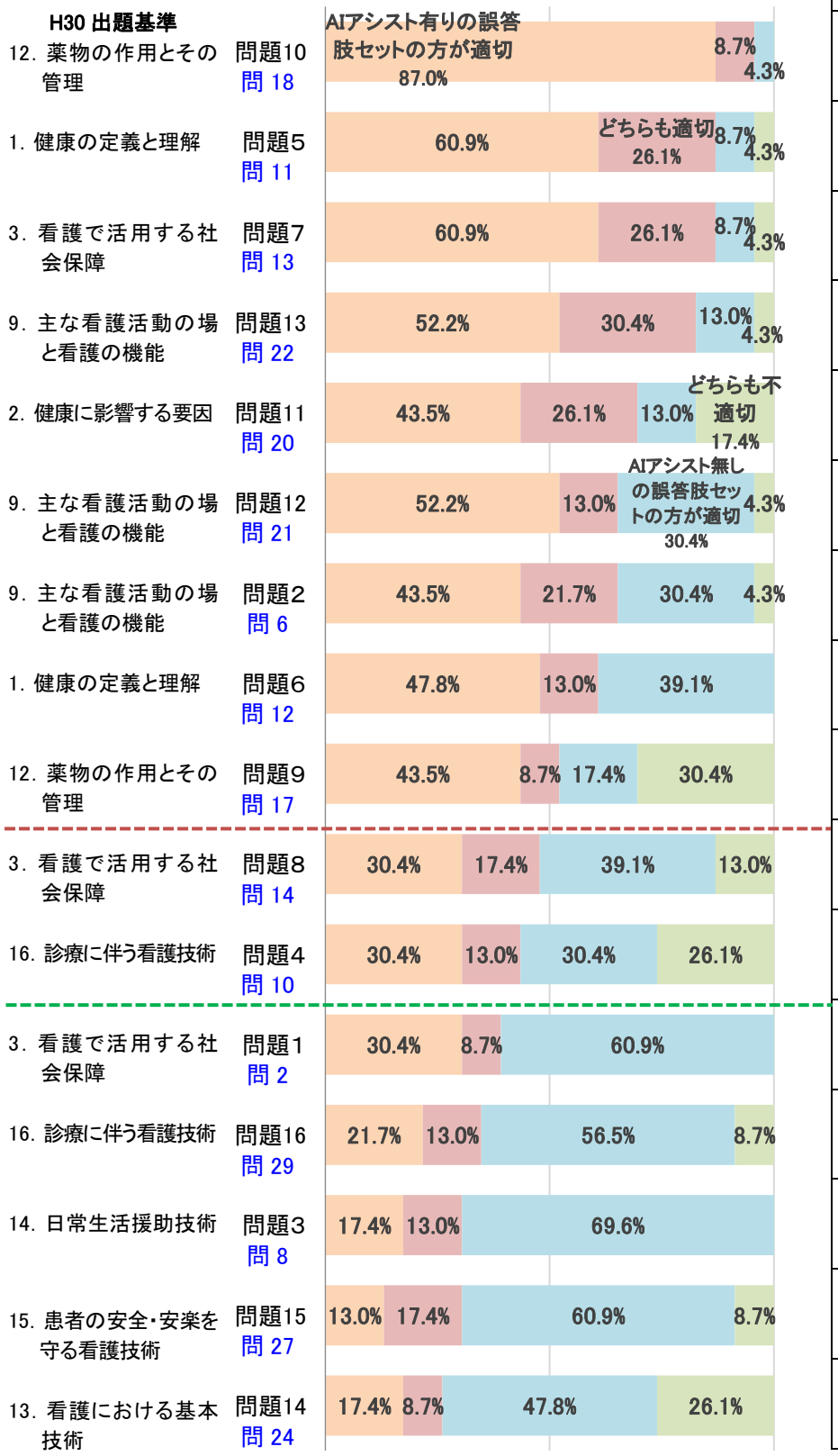
の方が適切、あるいはどちらのセットも適切と回答した人の合計の割合が「AI アシスト無し」の方が適切と判断した人の割合より高かった問題（図 13 の緑破線まで）は 16 問中 11 問（68.8%）であった。問題の内容をみると、大項目 12「薬物の作用とその管理」、同 1「健康の定義と理解」、同 9「主な看護活動の場と看護の機能」に関する問題は「AI アシスト有り」の誤答肢セットの評価が高く、同 14「日常生活援助技術」、同 15「患者の安全・安楽を守る看護技術」、同 16「診療に伴う看護技術」など、看護技術に関わる問題は「AI アシスト無し」の誤答肢セットが適切と判断される傾向がみられた。

次に、対象を教員経験年数の中央値(8.5 年)で経験の短い群（教員歴 8 年以下）と経験の長い群（9 年以上）の 2 群に分け、回答結果を比較した（図 14、図 15)。その結果、「AI アシスト有り」の誤答肢セットの方が適切と判断された問題、あるいは「どちらも適切」と判断された割合の高い問題は両群で大きな相違はなかったが、経験年数の短い群の方が、「AI アシスト有り」の誤答肢セットの方が適切、あるいは「どちらも適切」と評価する人の割合が経験年数の長い群と比べて 16 問中 10 問（62.5%）で高かった。また、「どちらも不適切」と評価した人がいた問題の数は、経験年数の短い群では 16 問中 6 問（37.5%）であったのに対し、経験年数の長い群では 10 問（62.5%）とその数が多かった。

また、各問題に対する学生の正答率、識別指数の結果とあわせて評価をしたところ、図 13 に示すように教員による誤答肢セットの評価と、学生の実際の解答状況との間に一定の傾向は見られなかった。

図 13　AI アシスト有無の各誤答肢セットの評価（「AI アシスト有りの方が適切＋どちらも適切」が多い順）

青字は学生対象調査結果

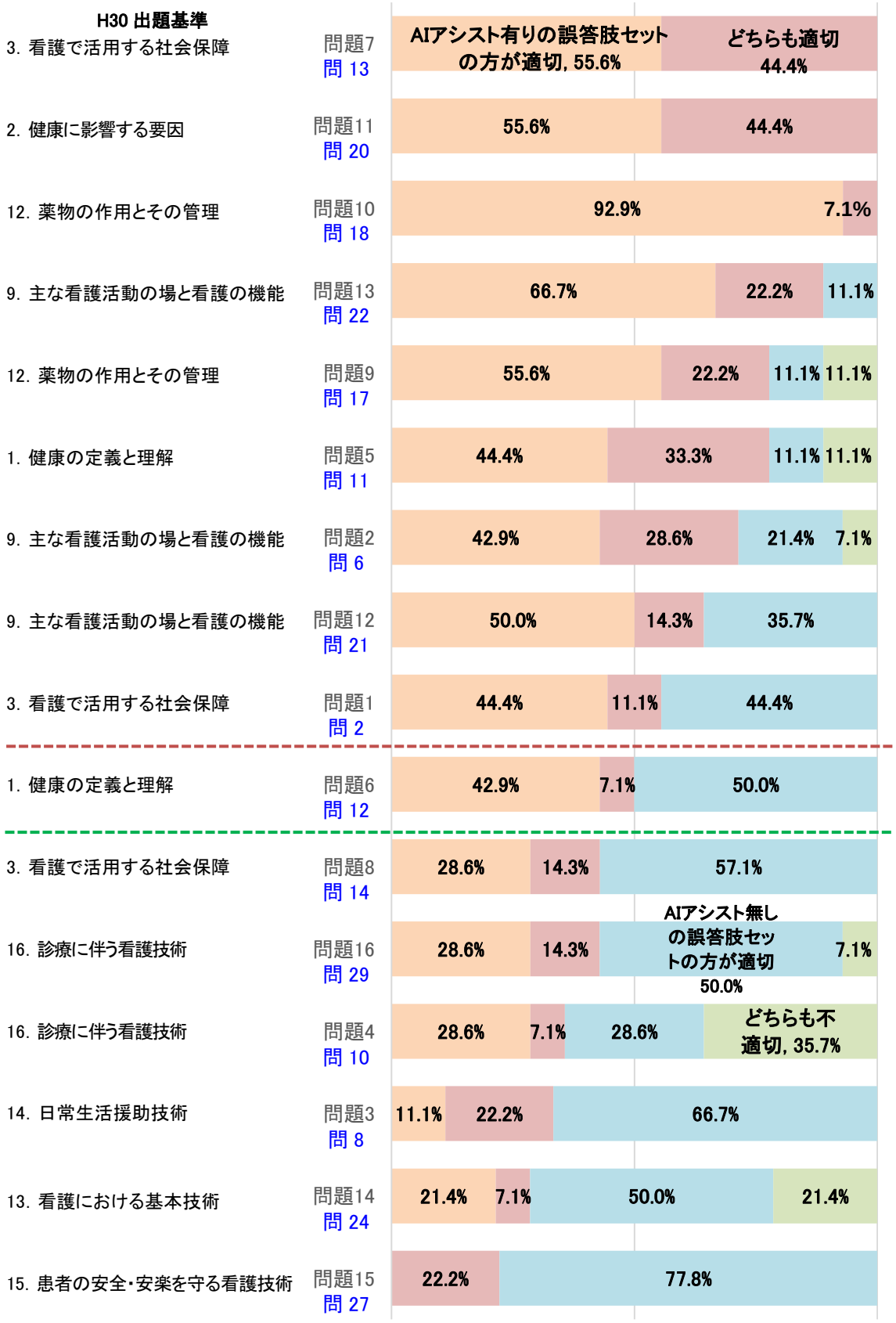


	正答率	識別指数	評価*
AI 有	84.5	0.447	②
AI 無	96.4	0.120	△
AI 有	95.2	0.085	△
AI 無	93.9	0.140	△
AI 有	94.1	0.128	△
AI 無	93.9	0.160	△
AI 有	64.2	0.574	②
AI 無	86.8	0.280	②
AI 有	98.9	0.021	△
AI 無	95.9	0.080	△
AI 有	98.4	0.021	△
AI 無	99.0	0.040	△
AI 有	64.2	0.468	②
AI 無	74.6	0.407	②
AI 有	96.3	0.149	△
AI 無	91.4	0.254	○
AI 有	81.8	0.468	②
AI 無	87.3	0.400	②
AI 有	96.8	0.064	△
AI 無	90.9	0.194	△
AI 有	85.6	0.362	②
AI 無	81.2	0.460	②
AI 有	69.0	0.511	②
AI 無	86.3	0.234	②
AI 有	89.3	0.234	②
AI 無	85.8	0.374	②
AI 有	63.6	0.532	②
AI 無	71.1	0.594	②
AI 有	79.1	0.553	②
AI 無	62.9	0.621	②
AI 有	97.3	0.064	△
AI 無	93.4	0.100	△

赤線：「AI アシスト有りの方が適切」＞「AI アシスト無しの方が適切」の境界
緑線：「AI アシスト有りの方が適切＋どちらも適切」＞「AI アシスト無しのほうが適切」の境界

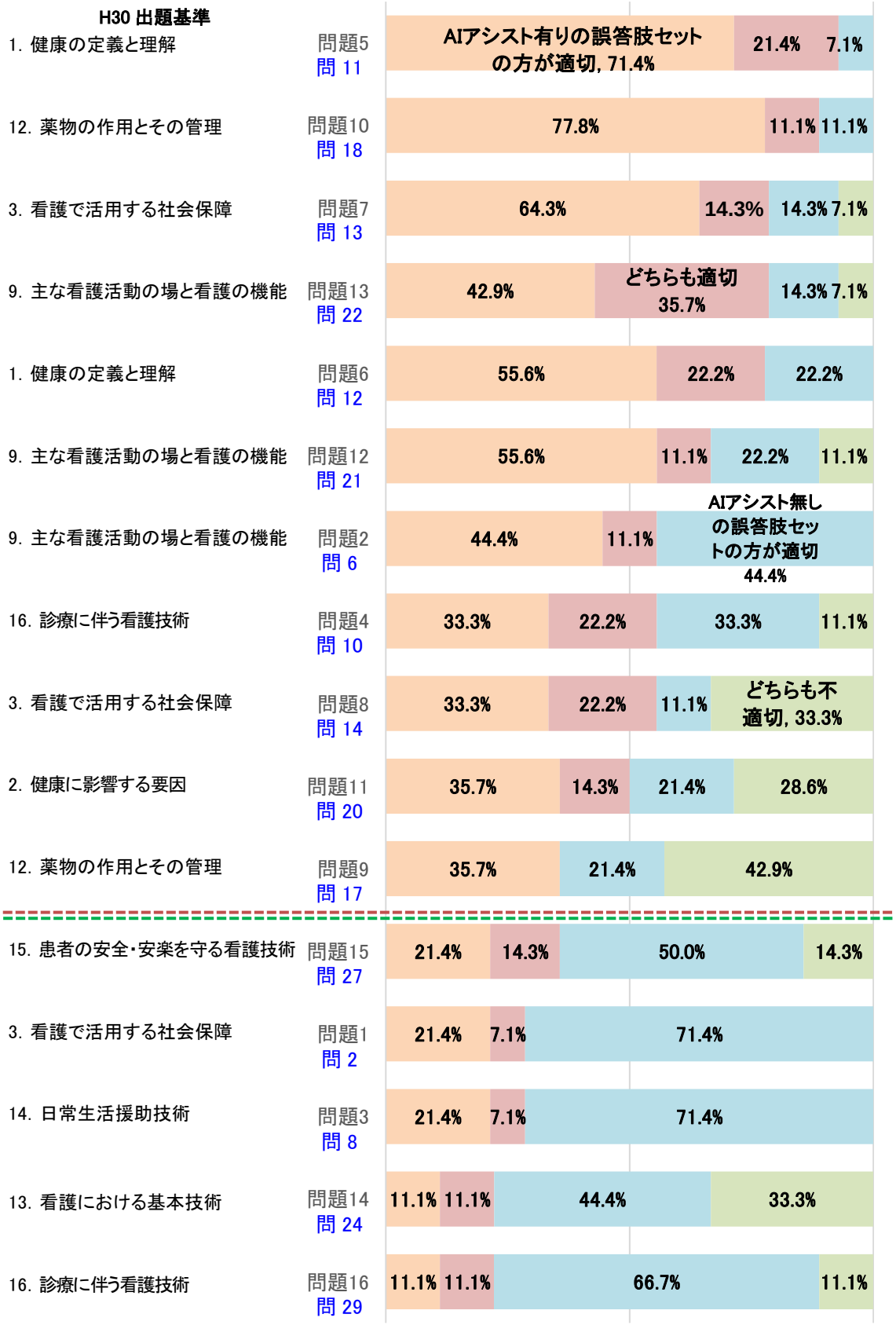
＊評価　p16 表 4 参照
○：良問　②：良問だが難問
△：非該当問題

図 14　AI アシスト有無の各誤答肢セットの評価　　看護教員経験年数 8 年以下



赤線：「AI アシスト有りの方が適切」＞「AI アシスト無しの方が適切」の境界
緑線：「AI アシスト有りの方が適切＋どちらも適切」＞「AI アシスト無しのほうが適切」の境界

図 15　AI アシスト有無の各誤答肢セットの評価　　看護教員経験年数 9 年以上



赤線：「AI アシスト有りの方が適切」＞「AI アシスト無しの方が適切」の境界
緑線：「AI アシスト有りの方が適切＋どちらも適切」＞「AI アシスト無しのほうが適切」の境界

D．考察

1.良問の基準について

看護師国家試験必修問題の良問ルールの抽出のために行った計量テキスト分析では、結果の共起ネットワークから良問に特徴的なルールを抽出することが困難であった。これは、必修問題が広範囲に渡り、各出題項目に関する問題数（テキストデータ）が限られるため、比較的抽象度の高い共起語が抽出されるに留まったことが要因として考えられた。

また、分析対象 500 問すべての正答率・識別指数をプロットし、先行研究[6]の結果と照合した結果、正答率 90～99%未満かつ識別指数 0.2 以上を『◯良問』とし、その他を①易しすぎる問題、②良問だが難問、③あまり適切ではない問題、△非該当問題（前述全てに当てはまらない問題）の 5 つに区分したことで、各問題の特性を明確にした。この基準に基づき、小項目、中項目ごとに類題を丹念に比較することで、良問とそれ以外の問題の相違点を見出すことが可能となったことから、本基準は国試必修問題の質評価の一指針となり、更に本作業を行うことでルールブックにおける基準の明確化に繋がったと考える。

2.良問作成のルールブックについて

問題分析については、初年度(令和 4 年度)に定めた問題評価の基準に則って過去問題を分類し、小項目単位での良問のルールを抽出することで、今後の国家試験作問に向けた作問ルールブックとして即利用可能な資料の作成に繋がると考える。また、作問ルールの文言を、大規模事前学習済み言語モデルのプロンプトに適用する前提で作成することで、作問システム開発の一助となると考える。

3.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 技術の活用策の検討：問題作成プロセス評価

1)誤答肢案の有効性について

3 年間の研究期間で、様々な LLM を検討し、最終的に 6 種類 7 モデルの LLM を使用して誤答肢案を作成した。採用された 48 の誤答肢案のうち、11(22%)は、誤答肢案を参照しない「AI アシスト無し群」(3G、4G）の作問者が作成した誤答肢と重複していたことから、これら 11 の誤答肢案は、看護師国試専門家が作成したものと同程度に高品質であると考えられる。さらに残りの誤答肢案は、「AI アシスト無し群」の作問者が思いつかないものであったことから、LLM が新規性の高い誤答肢を生成できる可能性が示唆された。

また、48 の誤答肢案と各問題で実際に予備校の模試で使用された誤答肢(真の誤答肢)の重なりは、48 中 6(13%)であり、これらの真の誤答肢案も「AI アシスト有り群」(1G、2G）の作問者に採用されていた。これより、LLM が生成した誤答肢案が、その品質と新規性の点で問題作成を支援できることが示唆された。なお、QS1（問 1-問 5）で LLM 作成の誤答肢案の採用率が高いのは、QS2（問 6-問 10）と比べて良問が多い大項目を含む、問題の偏りが一因と考えられた。

LLM の相違についてみると、使用した中では古いモデルである Swallow-2 が最も高い採用率を示した。一方、医学テキストで学習した MedSwallow-3 は、そのベースモデルの Swallow-3 よりも採用率が低い結果であった。この結果は、医療領域と看護領域の専門用語の違いが影響している可能性が考えられた。GPT-3.5 ファミリーを除くモデルは「真の」誤答肢をよく再現していた。しかし、高い再現率は必ずしも高い採用率につながらないことから、生成された誤答肢を評価するためには、あらかじめ用意した「真の」誤答肢をいかに再現できるかという観点だけでは、

不十分であることが示唆された。

誤答肢の種類による採用率の相違を見ると、名詞句の選択肢では 0.61、文の選択肢では 0.32 となった。データサイズが小さいため、決定的な結論を出すことはできないが、名詞句候補が文候補よりも採用される傾向があることが示唆された。

本調査結果から、誤答肢案の提示が作問の障害にはならないことが示された。問題セット別に見ると、QS1 では AI アシスト有り群の方がわずかではあるが、より多くの時間を作問に使っていたが(5.9 分と 4.8 分)、 QS2 ではそれぞれ 8.0 分、37.3 分と、AI アシスト有り群の方がはるかに短時間で作成していることが示され、全体平均の時間短縮は主に QS2 によるものであることが明らかとなった。この違いは、誤答肢案の採用率の分析と同様に、問題セットが含む大項目の違いが理由であると考えられ、要改善問題の多い大項目の問題セットの作問において、「AI アシスト有り」の方が、作問時間が短縮される可能性が示唆された。

2)誤答肢案を使用する作問者の特性について

LLM が生成した誤答肢案を採用しやすい作問者の特徴を検討した結果、明らかな特徴として、国家試験委員経験年数の相違が挙げられた。経験年数が長い群は LLM が生成した誤答肢案の採用率が低い傾向があった。経験年数の短い群は、自身の判断より、LLM の提案を受け入れやすいことが一因と考えられる。また、両群の問題セットの相違による影響を検討したところ、経験の短い群では、QS1(問 1～問 5 ）の採用率 1.0、QS2（問 6～問 10）の採用率 0.40 であった。一方、経験の長い群では各々0.53 と 0.43 であったことから、この点においても経験の短い群が質問セットの違い、すなわち良問が多い大項目の問題セット（QS1）と要改善問題の多い大項目の問題セット（QS2）の違いによる、AI アシストの誤答肢活用状況への影響を受けており、良問の作りやすい領域でより AI アシストの誤答肢を活用する傾向が示唆された。

また、作問時間についても経験の短い群では AI アシスト無し群が AI アシスト有り群と比べて 4.4 倍の時間を要していたのに対し、経験が長い群では両者の差はほとんどなかったことから、作問時間に関しても、経験の短い作問者にとり AI アシストは有効であることが示唆された。

4.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

1)学生対象調査

研究 1 における学生対象調査では、大規模言語モデルにより作成した誤答肢案を参照して作問された問題（AI アシスト有り問題）と誤答肢案なしで作問された問題（AI アシスト無し問題）の正答率等の特性を比較することを目的として、模擬試験形式の調査を行った。

作問時の誤答肢案の有無では、全体の得点率には差がなく、誤答肢セットの作成において LLM 活用が実用可能であることが示唆された。個別の問題では、誤答肢案の有無で正答率に差がある問題が認められたものの、設定された誤答肢に大きな問題は認められなかったこともあり、許容範囲内であると考えられる。

2）教員対象調査

LLM を看護師国家試験の誤答肢作成に活用した結果、約 7 割（16 問中 11 問）の問題で、「AI アシスト有り」で作成した誤答肢セットは従来型の人手による作問、すなわち「AI アシスト無し」で作成した誤答肢セットと遜色がないことが明らかとなった。とくに、「薬理作用」「健康の定義と理解」「看護かつどの場と看

護の機能」の大項目に属する問題においては、LLM の活用が有効であることが示唆された。一方で、「社会保障」「看護技術」に関わる問題においては、現時点での LLM の活用が必ずしも有効ではない可能性も示唆された。今後は、LLM 活用が必ずしも有効ではなかった領域の問題について、その原因を探索し、その解決のための LLM の洗練が課題である。

また、本調査の結果より、AI アシストによる誤答肢セットは、教員経験年数の短い対象の思考に近いことが示唆された。学生対象調査の結果から、AI アシスト有りの誤答肢セットは AI アシスト無しの誤答肢セットと遜色がないことが示されたが、誤答肢案の多様性を増やすためにも、教員経験年数の長い教員の思考も反映した誤答肢を作成できるよう LLM を改善すること、さらに LLM 活用の適用範囲を拡大し、作問の質の担保を図ることが課題である。

２.【研究２】看護師等国家試験への CBT 実装の可能性の検討

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、三浦、木村

A．研究目的

研究 2 では、看護師等国家試験におけるコンピュータを活用した試験（CBT：Computer-based Testing）の実装に向けて、CBT システムの試用版を開発・試行し、試験問題の妥当性について難易度、識別指数、IRT スコア等から評価すると共に、出題形式、問題管理システム、受験者側の受容性に関する調査を行い CBT 導入に向けた課題を明示することを目的とした。今年度は、看護学生を対象に、看護師国家試験必修問題の模擬試験問題を CBT と従来型の筆記試験の 2 種類の形式で実施し、試験形式の違いによる解答の差、疲労感の相違、回答のしやすさ等について調査するとともに、CBT に対する受容性を調査することを目的とした。

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.CBT システムに取り込む問題プールの作成と問題取り込み

研究分担者・協力者をコアメンバーとして、作問チームを結成した。国家試験等の過去問、看護系大学共用試験 CBT の実用化と教育カリキュラムに関する先行研究も参照し、画像や音声も積極的に取り入れて問題プールを構築することとした。CBT システムとして TAO を用い、本システムに看護国試の模擬試験問題の必修問題を令和 4 年度に 250 問。令和 5 年度に 100 問の取り込みを行った。

2.CBT 実装可能性に関するインタビュー調査

1)調査対象

研究分担者・研究協力者の機縁により、日本国内の看護系大学 1 校、専門学校 2 校に協力を依頼し、各校で学生 20 名程度と教員 5 名程度の参加者を募集した。

学生は 4 年課程の学校においては 3 年生以上、3 年課程 2 年生以上を対象とした。

2)調査方法

①学生対象インタビュー調査

自記式質問紙により、学年等の基本属性、学校内外での CBT 受験経験、パソコン等の情報機器の所持・使用状況を回答してもらった。次に、サンプル試験 1 として CBT システムの TAO にアクセスしてもらい、動画や音声を使用したマルチメディア問題 5 問、従来型の看護師国家試験の問題 5 問、合計 10 問を制限時間 15 分で解答してもらった。その後、サンプル試験 2 として、サンプル試験 1 のマルチメディア問題と同等の内容の問題 5 問、サンプル試験 1 で出題した従来型の看護師国家試験の問題と同じ問題 5 問、合計 10 問を筆記試験形式で制限時間 15 分で解答してもらった。

その後、5 名以内のグループを設定し、フォーカスグループインタビューを行った。インタビューでは、主に以下の内容を尋ねた。

- TAO 上で解答してもらったマルチメディア問題と、同等の内容の問題の解答しやすさや難易度等の感想

- TAO での問題への解答のしやすさ、筆記試験（紙ベースの試験）との比較
- マルチメディア問題を含めた CBT 形式での試験に向けた学習・試験対策に必要な教育内容やそれが現状の教育で提供されているか

②教員対象インタビュー調査

自記式質問紙により、年齢、性別等の基本属性、専門分野、教員および看護職の経験年数、国家試験対策の担当の状況に付いて回答してもらった。

次に、個別インタビューまたはグループインタビューを行った。インタビュー調査は、事前に学生を対象とした調査のサンプル試験 1 の問題を TAO 上で閲覧してもらったうえで実施した。インタビューにおける質問内容は主に以下のようなものであった。

- 看護師等国家試験が CBT で実施されることに対する賛否とその理由
- 問題を再利用するために過去問題を非公開とすることについての賛否、理由、非公開となった場合に試験対策や教育のために必要な情報・支援
- サンプル試験 1 のマルチメディア問題の難易度、試験問題としての適切性等についての感想、試験対策するにあたって必要な教育や支援

3)分析方法

学生に対する調査で解答してもらった、サンプル試験の正答状況をサンプル試験 1 (CBT)、サンプル試験 2(紙)の対応する設問間で比較した。インタビュー内容については、質的内容分析を行った。

4)倫理的配慮

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：22 - AC103)。

3.CBT(Computer Based Testing)による試験実施の課題に関する調査

本調査は、研究 1 の「9.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および CBT(Computer Based Testing)による実施の課題に関する調査」と同時に行った。研究方法については、研究 1 の B. 研究方法（p11～13）を参照のこと。

C．研究結果

1.CBT システムの試験的運用に向けた問題プールの準備と問題の取り込み

看護師国家試験の模擬試験を実施している会社に依頼し、過去に実施された模擬試験問題、解答、解説を計 500 問収集した。また、CBT 実装可能性の検討に向けたインタビュー調査で使用するため、動画や音声を用いた問題を 5 問作成した。

本研究で用いる CBT システムとしては TAO を用いることとし、試用を開始した。①で収集した模擬試験問題のうち 250 問を取り込んだ他、インタビュー調査のサンプル問題として作成した動画や音声を用いた問題 5 問の取り込みも行った。TAO 上でのサンプル問題は図 16 のとおりである。

2.CBT 実装可能性の検討に向けたインタビュー調査

1）学生対象インタビュー調査

大学 1 校 13 名、専門学校 2 校 14 名の学生が調査に参加した。 対象者の学年は大学 4 年生が 13 名、専門学校 3 年生が 4 名、専門学校 2 年生が 10 名であった。学校内外でのコンピュータを利用した試験の受験経験がある者は

16 名(59.3%)、自分専用のパソコンを所持しているのは 22 名(81.5%)、パソコンの使用頻度が「毎日」の者は 16 名(59.3%)、自分専用のタブレット端末を所持しているのは 26 名(96.3%)であった。

表 22 にサンプル問題の正答状況を示した。全体として正答率に大きな違いはなかった。

図 16　TAO に取り込んだ動画サンプル問題の例

表 22 サンプル問題の正答状況

		CBT		紙	
	度数	正答率		正答率	P[a]
合計	27	0.741		0.752	0.611
問1	27	1		0.89	0.083
問2	27	0.89		0.93	0.317
問3	27	0.48		0.59	0.18
問4	27	0.7		0.67	0.739
問5	27	0.37		0.41	0.796
問6	27	0.89		0.89	b
問7	27	0.89		0.89	b
問8	27	0.78		0.81	0.317
問9	27	0.56		0.59	0.317
問10	27	0.85		0.85	b

巻末資料表 28 にマルチメディアを活用した問題に対する感想のまとめを示した。紙ベースの試験と比較して、「状況設定は文章よりも理解しやすい」一方で「必要な情報を得るのは文章題よりも時間がかかる」といった意見が挙げられた。また、マルチメディアを活用した問題が出題されることによる、学習や試験対策への影響としては、「臨床に活かせる形で学習できる」といったことが語られていた。

次に、CBT システムでの回答形式についての回答のまとめを巻末資料表 29 に示した。「解答のマークシートへの転記が不要なので解答に集中できる」「クリックで問題に飛べる」といった点がメリットに感じられた一方で、「画面を見ることによる疲労」「時間配分がたてにくい」「メモ機能がないと不便」といった点がデメリット・不便な点として挙げられた。また、解答の入力や問題送り等の基本的な操作については説明がなくても問題なく対応できた一方、動画・音声の視聴可能回数や使用できる機能の説明、チュートリアル等がないと不安であるとする指摘もあった。

最後に、マルチメディアを活用した問題が含まれる試験のための学習に必要なものとして、シミュレータ・マルチメディア教材による学習を増やすことや、実習・演習で実際に体験することなどがより必要である、といった意見が挙げられた。(巻末資料表 30)

2）教員対象インタビュー調査

教員を対象とした調査では、大学教員 8 名、専門学校教員 9 名から協力が得られた。教員経験は 9 ヶ月から 27 年、看護職の経験は 2 年 9 ヶ月から 35 年 3 ヶ月と幅があった。学生の国家試験対策担当をしている者は 6 名であった。

看護師等国家試験の CBT 化についての意見のまとめを巻末資料表 31 に示した。

「マルチメディアを活用した問題を設定できる」「試験の複数回実施のハードルが下がる」「採点等の利便性が上がる」といった利点があり、概ね CBT 化には賛成との意見が多かった。一方、「マルチメディア問題による情報量の増大の試験への影響」「作問負担の増大」「公平性の担保」「試験実施のためのインフラ整備」「実施費用の高額化」「問題漏えい等のセキュリティ」等の懸念も示されていた。

マルチメディアを活用した問題については（巻末資料表 32)、難易度は紙ベースの試験（筆記試験）と同等程度であるとの意見が多く、マルチメディアを活用することで能力を多角的に評価できるため、国家試験としてはより適切であるというような、肯定的な意見が多かった。一方で、解答時間や問題数は現行の紙ベースの試験と同じでは難しいであろうといった意見や、視覚・聴覚の多様性やジェスチャー等の解釈の違いなどから外国人受験者への配慮が必要であるといった課題が挙げられた。また、マルチメディアを活用した問題を作成する際の課題として、モデルとして出演してもらう人のプライバシーといった倫理的問題や、動画の解釈の多様性といったことが挙げられていた。

次に、試験問題プール制のための問題非公開化についてついては賛否がわかれた（巻末資料表 33)。「国際的には非公開が一般的」、「問題の質保証の体制が適切に整備されていれば問題ない」といった意見があった一方、「過去問対策も学習の一部である」ため公開が望ましい、「予備校等で出口調査等をして対策されるので機能しないのでは」といった意見もあった。また、非公開化時に必要となる情報・支援としては、移行初年度のサポートを手厚くすること、詳細な出題基準、サンプル問題、公式問題集のように出題される問題がイメージできるような情報公開が必要であるとの意見が出されていた。

現状の教育におけるマルチメディア・シミュレーション教材の活用状況については、学習管理システム(Learning Management System: LMS)、ビデオ会議システム、動画教材、シミュレータ、模擬電子カルテなどが活用されていた（巻末資料表 34)。LMS の機能として提供されている簡易 CBT 機能も授業各回の理解度確認等に利用されていたほか、新型コロナウイルス流行時には単位認定試験にも活用されていた。こうした状況から、看護師等国家試験が CBT 化され、マルチメディア問題が出題されるようになったとしても、現状の教育から大幅な変更は不要との意見が多かったが、マルチメディアを活用した問題への対応のためマルチメディア教材・シミュレーション教材の重要度がより増大することや、試験対策として CBT システムの使用方法についての一定のトレーニングは必要であろうとの意見が示された。

CBT に移行する場合の準備期間として、対象となる学生が入学する時点で周知されている必要があるとの意見が多かった（巻末資料表 35)。

3. CBT(Computer Based Testing)による試験実施の課題に関する調査

1）回答状況および参加者の特徴、2）対象者の特徴については、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のC. 研究結果 8.1)学生対象調査(p27)を参照のこと。

3）試験方式による共通問題正答数、疲労感、解答のしやすさの比較

表 23 に試験方式による共通問題正答数、疲労感、解答のしやすさの比較の結果を示した（回答詳細は巻末資料図 17 参照)。共通問題の正答数には有意な差は認められなかった。また、AI アシスト有り問題および無し問題の両群の解答比較（表 24)、および全問題の筆記、CBT の両群の解答比較(表 25)を行った結果、いずれについても両群間で解答状況に有意な差はみられなかった。

解答後の疲労感は、筆記試験群で有意に高かった（p＝0.005)。解答のしやすさに関する項目はすべて、筆記試験群が有意に肯定的な回答をしていた。

表 23　共通問題の得点、試験後の疲労感、解答形式に対する評価の試験形式による比較

	筆記試験群	**CBT 群**	
	n = 193	n = 191	**p-value**
共通問題の得点	19.00 (18.00, 20.00)	19.00 (18.00, 20.00)	>0.9
疲労感	3.50 (2.00, 5.00)	3.00 (1.00, 4.00)	0.005
無回答	1	2	
解答に集中できたか	4.00 (4.00, 4.00)	4.00 (3.00, 4.00)	0.001
どちらかといえば集中できなかった	1 (0.5%)	9 (4.8%)	
どちらかといえば集中できた	38 (20%)	56 (30%)	
集中できた	153 (80%)	124 (66%)	
無回答	1	2	
文字の読みやすさ	4.00 (4.00, 4.00)	4.00 (3.00, 4.00)	0.002
読みにくかった	0 (0%)	1 (0.5%)	
どちらかといえば読みにくかった	1 (0.5%)	3 (1.6%)	
どちらかといえば読みやすかった	28 (15%)	49 (26%)	
読みやすかった	164 (85%)	137 (72%)	
無回答	0	1	
時間配分のしやすさ	4.00 (4.00, 4.00)	4.00 (3.00, 4.00)	<0.001
しにくかった	0 (0%)	1 (0.5%)	
どちらかといえばしにくかった	2 (1.0%)	13 (6.8%)	
どちらかといえばしやすかった	28 (15%)	50 (26%)	
しやすかった	163 (84%)	126 (66%)	
無回答	0	1	
消去法のしやすさ	4.00 (3.00, 4.00)	3.00 (2.00, 4.00)	<0.001
しにくかった	1 (0.5%)	20 (11%)	
どちらかといえばしにくかった	11 (5.7%)	50 (26%)	
どちらかといえばしやすかった	50 (26%)	51 (27%)	
しやすかった	131 (68%)	69 (36%)	
無回答	0	1	
メモのとりやすさ	4.00 (4.00, 4.00)	2.00 (2.00, 4.00)	<0.001
しにくかった	0 (0%)	33 (17%)	
どちらかといえばしにくかった	2 (1.0%)	79 (42%)	
どちらかといえばしやすかった	36 (19%)	29 (15%)	
しやすかった	155 (80%)	49 (26%)	
無回答	0	1	
後で解答する問題のチェック	4.00 (4.00, 4.00)	4.00 (3.00, 4.00)	<0.001
しにくかった	0 (0%)	10 (5.3%)	
どちらかといえばしにくかった	2 (1.0%)	18 (9.5%)	
どちらかといえばしやすかった	35 (18%)	56 (29%)	
しやすかった	156 (81%)	106 (56%)	
無回答	0	1	
解答の見直しのしやすさ	4.00 (4.00, 4.00)	3.00 (2.00, 4.00)	<0.001
しにくかった	0 (0%)	20 (11%)	
どちらかといえばしにくかった	4 (2.1%)	29 (15%)	
どちらかといえばしやすかった	31 (16%)	54 (28%)	
しやすかった	158 (82%)	87 (46%)	
無回答	0	1	

表 24　筆記試験群と CBT 群の平均値の比較①　　AI アシスト有り・無し問題別

	AI アシスト有り問題			AI アシスト無し問題		
	A(筆記試験) n = 94	C(CBT) n = 93	p 値*	B(筆記試験) n = 99	D(CBT) n = 98	p 値*
AI アシスト有無比較問題 30 問（30 点満点） 平均点（SD）	24.9(3.93)	24.9(4.09)	0.994	25.5(4.03)	25.2(4.54)	0.915
共通問題 20 問（20 点満点） 平均点（SD）	18.4(2.05)	18.3(2.13)	0.855	18.2(2.39)	18.1(2.67)	0.853
合計 50 問（50 点満点） 平均点（SD）	43.3(5.61)	43.2(5.97)	0.871	43.7(6.12)	43.3(6.93)	0.994

*Mann-Whitney の U 検定

表 25　筆記試験群と CBT 群の平均値の比較②

	A(筆記試験) n = 193	C(CBT) n = 191	p 値*
AI アシスト有無比較問題 30 問（30 点満点） 平均点（SD）	25.2(3.99)	25.1(4.33)	0.964
共通問題 20 問（20 点満点） 平均点（SD）	18.3(2.23)	18.2(2.43)	0.997
合計 50 問（50 点満点） 平均点（SD）	43.5(5.88)	43.2(6.48)	0.874

*Mann-Whitney の U 検定

D．考察

1.看護師等国家試験の CBT 化に対する準備状況と課題

国内の大学、専門学校の学生、教員を対象に、看護師等国家試験の CBT 化に対する準備状況や課題について調査を行った。

看護師等国家試験の CBT 化に対する教員の意見としては、概ね肯定的であり、マルチメディアを活用した問題を使用できることでより適切に能力を評価できることや、試験を複数回実施できるようになることへの期待が多く示された。一方、試験の複数回実施の基盤となる問題プール作成のための過去問題非公開化に対しては不安や懸念が示された。現時点では日本における看護師等国家試験の対策には過去問題の演習が大きな役割を果たしており、CBT 化のメリットの享受の前提となる問題プールの構築に対する理解を得るとともに、過去問題の演習に替わる試験対策の方法を普及していく必要がある。

次に教育に関しては全体として、CBT 化されたとしても出題基準等に大きな変化がなければ、現状行われている教育を大きく変更する必要性などの影響は少ないと考えられた。しか

しながら、マルチメディアを活用した問題への対策にはシミュレーション教材や動画・音声を用いた教材を使用できる方が効果的な学習ができることから、教育機関の財政状況により、こうした教材の導入に差異がある場合には、一定の配慮・支援が必要である。

学生側もコロナ禍を経て推進されたオンライン教育、シミュレーション教育の結果、CBT の試験形式やマルチメディアを活用した問題に対応する基礎的能力は備わっていると考えられた。

令和 5 年度の調査では、機縁法により対象者を募ったため、準備状況の分布についての代表性には限界があるものの、CBT 化を進めるにあたって検討すべき事項については概ね抽出できたと考えられる。教育現場による準備状況については、どのような形式で試験が行われるか、ある程度具体的に提示されなければ正確な回答が難しいことから、今後具体的な実施方法の方向性ができ次第、本研究で示された検討事項等について量的な調査を行うことで、より正確に準備状況を把握できると考えられる。

2.CBT(Computer Based Testing)による試験実施の課題に関する調査

本調査では、筆記試験と CBT の解答形式間で問題への正答率や解答のしやすさを比較することを目的として、模擬試験形式の調査を行った。

筆記試験、CBT の試験形式間の比較では、試験の正答率には差がみられなかったものの、疲労感や解答のしやすさ等の解答形式への評価では差がみられた。

疲労感については、令和 5 年度の調査において、CBT 形式は画面を見ることによる疲労が懸念されていたが、本調査では CBT 群のほうが試験後の疲労感が低いという結果となった。これは、今回の調査で解答してもらったのは、単純想起型の問題で、令和 5 年度は動画等を視聴するマルチメディア問題であったことが影響したものと考えられる。単純想起型の問題では、マルチメディア問題と比較して画面を注視する時間が短いこと、マークシートでの解答と比較して、マーク位置のずれ等の解答の人為的なミスを気にする必要性が低いため、疲労感が低かったのではないかと考えられる。

次に、解答のしやすさ等の解答形式の評価では、「解答に集中できたか」「文字の読みやすさ」「時間配分のしやすさ」「消去法のしやすさ」「メモの取りやすさ」「後で解答する問題のチェック」「解答の見直しのしやすさ」のすべてで筆記試験のほうが肯定的な評価であった。本調査においては対象者の 3 割程度は CBT の経験がなかったことに加え、経験があったとしても CBT に慣れていた者は少なかったと考えられる。また、消去法やメモ、解答の見直し、時間配分については、問題冊子のほうが一覧性が高く、直接書き込めるため、筆記試験のほうがやりやすかったと考えられる。これらについては、CBT システムでも支援するような機能は含まれているため、実際に CBT で実施する際にはこれらの機能を使用可能とするとともに、受験者がシステムに習熟する機会を確保し、試験の内容とは関係ない要素が試験結果に影響しないよう十分な準備をすることが必要であることが示唆された。

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

初年度(令和 4 年度)は、過去 10 年間の看護師国家試験必修問題 500 問について、正答率・識別指数における一定の基準をもとに 5 区分の評価(良問、易しすぎる問題、良問だが難問、あまり適切でない問題、いずれにも該当しない問題）に分類し、国家試験問題の現状と各設問の課題を提示した。この結果を示したことは、これまでの看護師国家試験の定量的評価に繋がるものと考える。また、5 区分の評価を軸に、問題形式、解答形式の観点で出題基準の項目ごとの問題分析を行うことで、項目単位での良問のルールを抽出することが可能となり、今後の国家試験作問に向けた作問ルールブックとして即利用可能な資料として期待される。

また、既存の大規模事前学習済み言語モデルを試用して作問を試みたことで、今後看護に関わるデータを強化したデータベースを構築し、それを利用した作問システム開発への示唆を得ることができた。

2 年目(令和 5 年度)は、看護師国家試験の過去 10 年分の必修問題（500 問）の分析結果を統合した、「作問に向けた改善提案（ルールブック)」の作成と、3 種類の大規模言語モデル（ChatGPT3.5、ChatGPT4、JSLM）による 4 つの作問システム（gpt4、gpt3.5-few shot learning、gpt3.5-Fine Tuning、JSLM）を用いた試験的作問を行い、専門家による問題評価と各モデルの生成性能を評価するための自動評価指標の作成を試みた。その結果、258 の改善提案が提示され、提案内容のさらなる分析の結果、200 の小項目に対して 147 のルールが抽出された。また大規模言語モデルの活用可能性の検討の結果、gpt3.5-FT が他と比べて再現性、精度の評価指標値が良いことが示された。

最終年度(令和 6 年度)は、看護師国家試験の問題作成における誤答肢作成支援として、大規模言語モデル（LLM）の活用可能性を検討した。看護師国家試験委員経験者 15 人を対象に調査を行った結果、6 種類 7 モデルの LLM を用いて作成された誤答肢案 102 のうち 48（47%）が作問者（調査対象）により採用され、さらに採用はされなかったが妥当と判断されたものは 22 あったことから LLM が生成した誤答肢案の 69%（70/102）が妥当だと判断された。また採用された 48 の誤答肢案のうち、11(22%)は誤答肢案を参照しないで作問するグループ（AI アシスト無し群）が作成した回答と一致したことから、AI アシストによる誤答肢案の一部は看護師国試作問者の思考と同程度で誤答肢を生成していると考えられ、その質の高さが確認された。また、LLM による誤答肢は、作問者が思いつかない新規性のある選択肢を提供できる可能性も示唆された。作問者の属性では、国家試験委員の経験年数が短いほど LLM の提案を受け入れる傾向があり、特に良問の多い大項目の問題において、AI アシストによる誤答肢案が活用されやすいことが明らかとなった。また、AI 支援により特に要改善問題が多い大項目の問題の誤答肢作成において、作問時間が短縮され、効率化の効果も確認された。

次に、大規模言語モデル（LLM）を利用して作成した誤答肢案を参照して作問された問題（AI アシスト有り問題）と誤答肢案なしで作問された問題（AI アシスト無し問題）の正答率、識別指数、並びに問題の質を評価することを目的として、国内の大学、専門学校の看護学生 384 人を対象に模擬試験形式の調査を、看護教員 46 人を対象に質問紙調査を行った。

学生対象の模擬試験形式の調査では、AI 支援の有無による得点率に大きな差は見られず、LLM を活用した誤答肢作成が教育的視点から実用可能であることが示された。教員を対象と

した調査では、AI 支援によって作成された問題の多くが従来の作問と同等の質であると評価された。特に「薬理作用」や「看護活動の場と機能」など一部の分野においては、LLM の支援が有効であることが確認された一方、「社会保障」や「看護技術」に関しては、LLM の適用に限界がある可能性も指摘された。

以上の結果から、LLM は看護師国家試験の必修問題の誤答肢作成における支援ツールとして有用であることが明らかとなった。今後は、作問者の多様な思考を反映できるよう LLM の性能をさらに洗練させるとともに、適用範囲を拡大し、作問の質の担保を図ることが課題である。

2.【研究２】看護師等国家試験への CBT 実装の可能性の検討

初年度は、CBT システムの試験運用に向けた問題収集・作成と、CBT 実装可能性の検討に向けたインタビュー調査の準備を行った。試験運用に向けた問題は 500 問収集でき、動画や音声等のマルチメディアを活用した問題の作成にも着手した。インタビュー調査については，年度末に研究倫理審査委員会の承認を得て，新年度から調査開始の運びとなった。

2 年目は国内の大学、専門学校の学生、教員を対象に、看護師等国家試験の CBT 化に対する準備状況や課題について調査を行った。看護師等国家試験の CBT 化に対する教員の評価は概ね肯定的であったが、複数受験が可能になることなどの CBT 化のメリットを享受するために必要な問題プールの構築に関しては時間をかけて理解を得ていく必要があると考えられた。学生の CBT の受容性に関しては、コロナ禍を経て推進されたオンライン教育、シミュレーション教育の結果、CBT の試験形式やマルチメディアを活用した問題に対応する基礎的能力は備わっていると考えられた。

最終年度は、筆記試験と CBT の解答形式の違いによる正答率や解答のしやすさの相違を明らかにすることを目的として、国内の看護系大学、看護専門学校の学生 384 人を対象に模擬試験形式の調査を行った。

筆記試験、CBT の試験形式間の比較では、試験の正答率には差がみられなかったものの、疲労感は CBT 群の方が有意に低く（$p=0.005$）、解答に集中できたか（$p=0.001$）、文字の読みやすさ（$p=0.002$）、時間配分のしやすさ、消去法のしやすさ、メモの取りやすさなど、解答のしやすさについては、全ての項目において筆記試験群の方が肯定的な評価をしたものが多かった（$p<0.001$）。今後、看護師等国家試験を CBT で実施することを検討する際には、消去法やメモ等を筆記試験での解答と同様に実行できる機能を使用可能とするとともに、受験者がシステムに習熟する機会を確保し、試験の内容とは関係ない要素が試験結果に影響しないよう十分な準備をすることが必要であることが示唆された。

F．健康危険情報

該当なし

G．研究発表

1．論文発表

Yusei Kido, Hiroaki Yamada, Takenobu Tokunaga, Rika Kimura, Yuriko Miura, Yumi Sakyo and Naoko Hayashi, “Automatic Question Generation for the Japanese National Nursing Examination using Large Language Model”, Proceedings of the 16th International Conference on Computer Supported Education - Volume 1: pp.821-829,2024. DOI:10.5220/0012729200003693.

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 年

Yusei Kido, Hiroaki Yamada, Takenobu Tokunaga, Rika Kimura, Yuriko Miura, Yumi Sakyo 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 - Volume 1: pp.754-764, 2025. DOI: 10.5220/0013460300003932.

２．学会発表

Yusei Kido, Hiroaki Yamada, Takenobu Tokunaga, Rika Kimura, Yuriko Miura, Yumi Sakyo and Naoko Hayashi, “Automatic Question Generation for the Japanese National Nursing Examination using Large Language Models”, The 16th International Conference on Computer Supported Education, Angers, 2024/6.

米倉佑貴、林 直子、木村理加、佐居由美、三浦友理子、佐伯由香、佐々木幾美、宮本千津子、「看護師等国家試験の Computer Based Testing(CBT)形式での実施可能性に関する調査」、第 44 回日本看護科学学会学術集会、熊本、2024 年 12 月.

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、長崎、2025 年 3 月.

Yusei Kido, Hiroaki Yamada, Takenobu Tokunaga, Rika Kimura, Yuriko Miura, Yumi Sakyo, Naoko Hayashi, “Evaluation of LLM-generated Distractors of Multiple-choice Questions for the Japanese National Nursing Examination”, The 17th International Conference on Computer Supported Education, Porto, 2025/4.

H．知的財産権の出願・登録状況

該当なし

引用文献

1) Sheng Shen, Yaliang Li, Nan Du, Xian Wu,Yusheng Xie, Shen Ge, Tao Yang, Kai Wang, Xingzheng Liang, Wei Fan(2020). On the Generation of Medical Question-Answer Pairs, ArXiv,abs/1811.00681.
2) Mark J GierlHollis LaiHollis LaiSimon R Turner(2012). Using automatic item generation to create multiple-choice test items, Medical Education, 46(8), 757-65.
3) Mark J. Gierl, Hollis Lai(2018). Using Automatic Item Generation to Create Solutions and Rationales for Computerized Formative Testing, Applied Psychological Measurement, 42(1), 42-57.
4) Filipe Falcão, Patrício Costa & José M. Pêgo(2022). Feasibility assurance: a review of automatic item generation in medical assessment, Advances in Health Sciences Education,https://doi.org/10.1007/s10459-022-10092-z.
5) Dhawaleswar Rao CH and Sujan Kumar Saha(2020). Automatic Multiple Choice Question Generation From Text :A Survey,

IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES,13(1),14-25
6) 令和元年度厚生労働科学研究「保健師助産師看護師国家試験における現状の評価及び出題形式等の改善に関する研究林」総括・分担研究報告書、林直子（研究代表者）他、2020 年 3 月.
7)D. R. CH and S. K. Saha. Automatic multiple choice question generation from text : A survey. IEEE Transactions on Learning Technologies, 13(1):14--25, 2020.
8) C.-Y. Lu and S.-E. Lu. A survey of approaches to automatic question generation: from 2019 to early 2021. In The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021), pages 151-162, 2021.
9) G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari. A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education, 20:121-204, 2020.
10) L. Pan, W. Lei, T.-S. Chua, and M.-Y. Kan. Recent advances in neural question generation. arXiv:1905.08949, https://doi.org/10.48550/arXiv.1905.08949, 2019.

表 26　「問題作成プロセス評価」　各問題回答詳細

問題1	令和元年(2019 年)の主要死因別にみた死亡率が最も高いのはどれか。						
正答	悪性新生物						
AI 生成 誤答肢	①心疾患　②脳血管疾患　③肺炎 ④老衰　⑤循環器系疾患　⑥消化器系疾患						
平均値	作問時間（分）	負担感	参考度	ブレスト貢献	阻害度	有効・採用誤答肢数	
AI アシスト有り	4.0	1.0	4.3	4.5	1.8	3.0	①心疾患　②脳血管疾患 ③肺炎　④老衰
AI アシスト無し	2.8	2.0	−	−	−	−	−

問題2	介護保険による給付はどれか。						
正答	地域密着型サービスの費用						
AI 生成 誤答肢	①医療機関の建設費用　②医療保険の自己負担分 ③入院治療の費用　④医療機関の入院費用 ⑤介護福祉士の資格取得の費用　⑥介護老人保健施設の費用 ⑦介護保険施設の入所費用　⑧住宅ローンの返済 ⑨健康診断の費用　⑩診療報酬 ⑪介護施設の入所費用　⑫介護老人福祉施設の建設費 ⑬介護療養型医療施設の費用　⑭介護保険施設の食費						
平均値	作問時間（分）	負担感	参考度	ブレスト貢献	阻害度	有効・採用誤答肢数	
AI アシスト有り	9.5	2.5	3.8	3.8	2.3	2.8	②医療保険の自己負担分 ③入院治療の費用 ④医療機関の入院費用 ⑦介護保険施設の入所費用 ⑨健康診断の費用 ⑪介護施設の入所費用 ⑭介護保険施設の食費
AI アシスト無し	8.3	3.0	−	−	−	−	−

問題3	臓器の移植に関する法律における脳死の判定基準に該当するのはどれか。						
正答	平坦脳波						
AI 生成 誤答肢	①瞳孔散大　②高血圧　③心拍停止 ④心電図の平坦化　⑤心拍数の増加　⑥徐脈 ⑦無痛覚　⑧意識喪失　⑨心停止 ⑩血圧の低下　⑪自発呼吸						
平均値	作問時間（分）	負担感	参考度	ブレスト貢献	阻害度	有効・採用誤答肢数	
AI アシスト有り	5.0	2.3	4.5	4.3	1.5	2.5	①瞳孔散大　③心拍停止 ④心電図の平坦化 ⑦無痛覚　⑧意識喪失 ⑨心停止　⑪自発呼吸
AI アシスト無し	3.5	2.8	−	−	−	−	−

問題4	副腎皮質ステロイドの副作用〈有害事象〉はどれか。						
正答	易感染						
AI 生成誤答肢	①高血圧 ④食欲減退 ⑦視力向上		②血圧上昇 ⑤多尿 ⑧血圧低下		③体重減少 ⑥体重増加 ⑨多汗		⑩低血糖
平均値	作問時間（分）	負担感	参考度	ブレスト貢献	阻害度	有効・採用誤答肢数	
AI アシスト有り	5.5	2.0	4.5	4.3	1.3	3.3	①高血圧 ②血圧上昇 ④食欲減退 ⑤多尿 ⑥体重増加 ⑧血圧低下 ⑨多汗 ⑩低血糖
AI アシスト無し	3.5	2.5	−	−	−	−	−

問題5	大気汚染物質はどれか。						
正答	光化学オキシダント						
AI 生成誤答肢	①二酸化炭素 ⑤オゾン層 ⑨トリクロロエチレン		②水銀 ⑥水蒸気 ⑩二酸化硫黄		③カドミウム ⑦紫外線		④硫黄酸化物 ⑧窒素酸化物
平均値	作問時間（分）	負担感	参考度	ブレスト貢献	阻害度	有効・採用誤答肢数	
AI アシスト有り	5.5	2.3	4.8	4.5	1.8	3.8	①二酸化炭素 ②水銀 ③カドミウム ④硫黄酸化物 ⑧窒素酸化物 ⑨トリクロロエチレン ⑩二酸化硫黄
AI アシスト無し	5.8	2.0	−	−	−	−	−

問題6	保健所の役割はどれか。						
正答	感染症対策						
AI 生成誤答肢	①医療費の支給 ④診療 ⑦交通規制 ⑩予防接種 ⑬難病の治療		②道路整備 ⑤災害対策 ⑧交通事故の調査 ⑪介護認定の審査		③教育カリキュラムの作成 ⑥生活習慣病の予防 ⑨療養 ⑫医療機関の開設許可		
平均値	作問時間（分）	負担感	参考度	ブレスト貢献	阻害度	有効・採用誤答肢数	
AI アシスト有り	8.3	1.8	3.8	2.8	2.8	2.3	⑤災害対策 ⑥生活習慣病の予防 ⑩予防接種 ⑪介護認定の審査 ⑫医療機関の開設許可
AI アシスト無し	28.3	2.0	−	−	−	−	−

問題7	関節可動域を表すのはどれか。						
正答	ROM						
AI 生成 誤答肢	①BMI ⑥BP	②BMR ⑦ROMI	③MRI ⑧AED	④ECG ⑨ROMII	⑤HR		
平均値	作問時間（分）	負担感	参考度	ブレスト貢献	阻害度	有効・採用誤答肢数	
AI アシスト有り	6.0	2.5	1.8	1.5	2.8	0.3	②BMR
AI アシスト無し	33.3	3.0	−	−	−	−	−

問題8	成人男性への持続的導尿（膀胱留置カテーテル挿入）で正しいのはどれか。						
正答	カテーテルの固定は下腹部に行う。						
AI 生成 誤答肢	①カテーテルの固定は胸部に行う。 ③カテーテルの固定は太ももに行う。 ⑤カテーテルは陰茎の根元から挿入する。 ⑦カテーテルの固定は足首に行う。 ⑨カテーテルは陰茎の先端から挿入する。 ⑪カテーテルの固定は鼠径部に行う。			②カテーテルの固定は大腿部に行う。 ④カテーテルの固定は膝上に行う。 ⑥カテーテルの挿入は無菌操作を必要としない。 ⑧カテーテルの先端は洗面器に浸す。 ⑩カテーテルの固定は上腹部に行う。			
平均値	作問時間（分）	負担感	参考度	ブレスト貢献	阻害度	有効・採用誤答肢数	
AI アシスト有り	8.5	2.5	2.5	1.5	2.3	1.0	②カテーテルの固定は大腿部に行う。 ⑩カテーテルの固定は上腹部に行う。 ⑪カテーテルの固定は鼠径部に行う。
AI アシスト無し	40.0	2.7	−	−	−	−	−

問題9	標準予防策（スタンダードプリコーション）で感染源として取り扱うのはどれか。						
正答	尿						
AI 生成 誤答肢	①汗	②唾液	③涙	④ 血液	⑤鼻水		
平均値	作問時間（分）	負担感	参考度	ブレスト貢献	阻害度	有効・採用誤答肢数	
AI アシスト有り	6.0	3.0	2.3	2.3	1.5	1.8	①汗　②唾液 ③涙　④血液 ⑤鼻水
AI アシスト無し	38.3	3.0	−	−	−	−	−

問題 10	成人の経鼻経管栄養法で正しいのはどれか。
正答	カテーテルの位置確認のために胃液を吸引する。
AI 生成	①カテーテルの位置確認のために空気を注入する。 ②カテーテルの挿入は患者が仰向けの状態で行う。

誤答肢	③カテーテルの位置確認は必要ない。 ④経鼻経管栄養は食事の代替として行われる。 ⑤カテーテルは鼻孔から挿入する。 ⑥栄養剤は常温で使用する。 ⑦カテーテルの挿入後、すぐに栄養剤を注入する。 ⑧カテーテルの位置確認のために血液を吸引する。 ⑨カテーテルの位置確認のために生理食塩水を注入する。 ⑩経鼻経管栄養は胃にカテーテルを挿入する。 ⑪カテーテルは口腔から挿入する。 ⑫栄養剤の投与速度は速やかに行う。 ⑬カテーテルの位置確認のために水を注入する。

平均値	作問時間（分）	負担感	参考度	ブレスト貢献	阻害度	有効・採用誤答肢 数	有効・採用誤答肢
AI アシスト有り	12.0	2.5	3.5	2.0	2.3	1.8	②カテーテルの挿入は患者が仰向けの状態で行う。 ⑥栄養剤は常温で使用する。 ⑦カテーテルの挿入後、すぐに栄養剤を注入する。 ⑨カテーテルの位置確認のために生理食塩水を注入する。 ⑬カテーテルの位置確認のために水を注入する。
AI アシスト無し	46.7	2.7	－	－	－	－	－

表 27　教員対象調査　各問題回答詳細

<table>
<tr><td>問題1</td><td colspan="2">介護保険による給付はどれか。</td><td>正答選択肢：地域密着型サービスの費用</td></tr>
<tr><td>誤答肢セット1</td><td>AI アシスト有り誤答肢
◆：AI 作成誤答肢</td><td colspan="2">医療機関の入院費用◆
健康診断の費用◆
介護保険施設の食費◆</td></tr>
<tr><td>誤答肢セット2</td><td>AI アシスト無し誤答肢</td><td colspan="2">介護予防サービスの費用
家事代行サービスの費用
訪問理髪サービス</td></tr>
<tr><td>誤答肢セットの適切性評価</td><td colspan="3"> 
</td></tr>
<tr><td colspan="4">誤答肢評価の理由</td></tr>
<tr><td>誤答肢セット1(AI アシスト有)が適切</td><td colspan="3">• 介護保険の給付内容がわかっていないと回答できないと考えたから。誤答肢2はサービス内容が詳細で選びにくいと感じる。
• 誤答肢2の中に正答があると、地域密着型サービスに対して、他の選択肢が小さく見える
• 誤答肢1はどの選択肢も正答と抽象度が近いような気がしました。誤答肢2の「訪問理髪サービス」は「の費用」まで入れたほうが良いかと思いました。
• 回答の抽象度が正答と合ってると思ったため
• 幅開く読み取れるため
• 看護に関連する選択肢のため。正答の地域密着型サービスの費用は介護保険における介護給付であり、誤答肢2の介護予防サービスの費用は介護保険における予防給付のため、どちらも介護保険における給付になるため、誤答とはできないのではないか。</td></tr>
<tr><td>誤答肢セット2(AI アシスト無)が適切</td><td colspan="3">• どの回答も当てはまると思ってしまう
• 正しい知識がないと、介護保険という文字を見て正解だと考えると思う
• サービスの費用という語句が同じなので、見間違いや戸惑う可能性があるから。
• 介護に限定されていて迷いやすい
• 誤答肢1には医療機関の入院費用や健康診断の費用など明らかに誤答であると分かりやすいものが入っている。誤答肢（AI無）の方には正答選択肢と同じ「…サービスの費用」という表現で示されており、迷いやすい。
• 誤答肢2は各選択肢にサービスという言葉が使われており、介護保険で給付を受けることのできるサービスについて整理できていないと回答しづらいと考えるため。
• サービスで統一され、訪問などの記載があると迷うと思うため
• 地域の社会資源を活用したサービスを想起させる選択肢であるから
• 地域密着型ということから介護、訪問などのキーワードがあるため
• 正答選択肢の「～サービスの費用」と似たような選択肢表現になることで回答者は惑わされる。
• 誤答肢2は在宅で使用するイメージがあるため、誤答肢1より魅惑肢になると思う。誤答肢2の選択肢が訪問理髪サービス(の費用)、となっていないのが気になりますが、全ての選択肢がサービスで統一されているところも良いと思う。</td></tr>
<tr><td>どちらも適切</td><td colspan="3">• 明らかに正解があるから
• 不適切な回答ないため</td></tr>
<tr><td>どちらも不適切</td><td colspan="3">（回答なし）</td></tr>
</table>

問題2	保健所の役割はどれか。	正答選択肢：感染症対策
誤答肢セット1	AI アシスト有り誤答肢 ◆：AI 作成誤答肢	介護認定の審査◆ 生活習慣病の予防◆ 精神疾患の予防
誤答肢セット2	AI アシスト無し誤答肢	地域ケア会議 母子健康手帳の交付 要介護認定
誤答肢セットの適切性評価		
誤答肢評価の理由		
誤答肢セット1(AIアシスト有)が適切	• 誤答肢2よりも間違いやすい問題となっている。 • 誤答肢2は介護保険関係が 2 つ入っているため • 母子健康手帳や要介護認定は、当校学生の正答率は悪くないから • 生活習慣病の予防という選択肢が保健活動として適切そうであった • 内容の範囲の抽象度が誤答肢 1 の方が合ってあるように感じたため • 保健という点で「予防」という視点が学生にとっては強いものだと考えたから • 誤答肢1については、予防というワードが魅惑肢となる可能性があるが、誤答肢 2 に関しては迷うものがないため不適切だと考える • 保健所と保健師で誤解されやすいか	
誤答肢セット2(AIアシスト無)が適切	• 各選択肢の内容が明確であるため • 誤答肢セット 1、2 は不正解と言えるのか疑問に感じたため • 母子保健については保健所業務と根付いている気がしました。 • 1 は、生活習慣病の予防や精神疾患の予防と漠然としていている選択肢と感じる	
どちらも適切	• 誤答肢1：保健師と保健所を読み違えていると生活習慣病の予防にとしてしまいそう。誤答肢2：保健センターと混同しそう（母子健康手帳の交付） • 誤答肢1はキーワードだけを覚えている学生が間違えやすく、誤答肢 2 は「なんとなく」の雰囲気で覚えていると間違えると思ったため • 介護領域の「審査や認定」は誤答肢であると分かりやすいが、その他の選択肢はいずれも公衆衛生 に関する業務に思えたため。	
どちらも不適切	• 保健所の業務を問うため	

問題 3	成人男性への持続的導尿（膀胱留置カテーテル挿入）で正しいのはどれか。	正答選択肢：カテーテルの固定は下腹部に行う。
誤答肢セット 1	AI アシスト有り 誤答肢 ◆：AI 作成誤答肢	カテーテルの固定は上腹部に行う。◆ カテーテルの固定は鼠径部に行う。◆ カテーテルの固定は大腿部に行う。◆
誤答肢セット 2	AI アシスト無し 誤答肢	陰茎を 60 度の角度にしてカテーテルを挿入する。 カテーテルの挿入は外尿道口から 4～6cm を目安とする。 固定用バルーンには生理食塩水を入れる。
誤答肢セットの適切性評価		
誤答肢評価の理由		
誤答肢セット1（AI アシスト有）が適切	• わかりやすい • 上腹部ということで腹部の上とイメージするため	
誤答肢セット2（AI アシスト無）が適切	• 誤答肢 1 は、固定部位さえ覚えていれば答えられるが、誤答肢 2 は、すべての知識を覚えていないと答えられないため、迷う • 誤答肢 1 は、臨床および複数の教科書で違う視点があるため • カテーテル挿入に関することを尋ねてあるため、誤答肢 2 の方が選択肢として妥当であると思うため。 • 導尿に関して、固定以外も知識として身につけて欲しい • カテーテルに関する知識がないと選択をしぼれないから。 • 誤答肢 1 は、カテーテルの固定の根拠はさまざまあり（長期間留置なのか短期間留置なのかなど）、回答者に知識と根拠があっての迷いやすいと考えた。誤答肢 2 は、基礎的知識を問うものではあるが、正しい知識がないと回答できないと考える。 • 問題文が導尿と広い範囲であるため、解答も固定に限定しないほうがよい • 固定を問うのではなく、手技の根拠全部理解しておく必要があるので、誤答肢 2 の方が適切 • 誤答肢 2 の方が臨床上誤答であると認識しなければならない項目が多く含まれる。 • カテーテルの固定位置のみを聞くよりは、持続的導尿における他に必要な知識も尋ねた方が良いと感じました。 • 固定方法だけでなく、持続的導尿に必要な知識を問うているため • 場所のみの問いとなっていないから • 細かな知識が必要 • 正しい手技が問われている選択肢のようなので。 • 誤答肢 1 だと固定位置がわかっているかどうかしか確認できないが、誤答肢 2 だと手技の流れ全体の理解度や解剖生理の理解度も確認できると思うため。 • 下腹部にテープ固定をするのであって、固定をするわけではない。そしてなにより、陰茎を上向きにして下腹部にテープ固定することが大切なのであって、下腹部に固定しても、陰茎が上を向いていなければ意味がないため、不適切だと思った。	
どちらも適切	• どちらの誤答肢セットにおいても正答が分からないと回答できないと考えるため。 • どちらも魅惑肢であるが、誤答肢2の方が選択肢の種類が多く、知識を問う問題として適当ではないか	
どちらも不適切	（回答なし）	

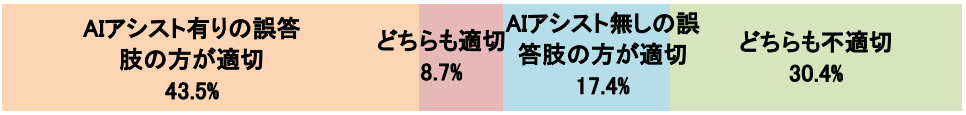
問題 4	成人の経鼻経管栄養法で正しいのはどれか。	正答選択肢：カテーテルの位置確認のために胃液を吸引する。
誤答肢セット 1	AI アシスト有り 誤答肢 ◆：AI 作成誤答肢	カテーテルの位置確認のために水を注入する。◆ カテーテルの挿入は患者を水平仰臥位にして行う。 経鼻経管栄養カテーテルは 30cm 程度を目安とする
誤答肢セット 2	AI アシスト無し 誤答肢	カテーテル挿入時は仰臥位とする。 カテーテルは鼻腔から 30cm 挿入して固定する。 カテーテルは毎日交換する。
誤答肢セットの適切性評価		
誤答肢評価の理由		
誤答肢セット1（AI アシスト有）が適切	• 毎日交換は学生は否定できると思うから • 栄養剤を注入するための管なので、何かを注入してみる（注入できる）ことが位置確認になると考える学生がいそう。水平仰臥位とは聞きなれない言葉なので、挿入のための特別な体位なのかもしれないと考えるかも。 • 正しい知識と比較しやすいため • 胃液確認について重要事項の知識を問うにあたって誤答肢 1 が適切であると感じた • 誤答肢 2 の毎日交換は実習をしている学生にとっては「魅惑肢」にはならないと思いました。誤答肢 1 であれば、水を注入している場面を見た経験がある学生にとっては魅惑肢になり得ると思いました。仰臥位も 30 センチのどちらも「魅惑」的ではないと思います。	
誤答肢セット2（AI アシスト無）が適切	• 水を注入するというのがリスクが高い • その都度交換すると思ってしまう可能性がある • ディスポーザブルに慣れている学生が多いため、「毎日交換」は惑わされやすい • 誤答肢 1 は日本語として不適切であり、カテーテルは 30cm 程度何の目安になるのかがわからない。水平仰臥位という表現も聞きなれない。	
どちらも適切	• どちらも迷う問題が組み込まれている。 • 困難さはない	
どちらも不適切	• 確認の際の体位やチューブは、毎日交換しない • 体位や長さなど明らかに不正確であり、魅惑肢にはならない • 誤答肢 1 は、3 番目が説明不足と感じた誤答肢 2 にある答えのほうが良いと感じた誤答肢 2 は、1 番目が説明不足と感じた誤答肢 1 にある答えの方が良いと感じた〃 • 長文である誤答肢 1 の方が読み取りに時間がかかるので魅惑肢になるかと思ったが、正答選択肢と同様な文章があるため、正答がわからずとも感覚で 2 択にするのではないかと考える。 • 誤答の内容が容易に誤答だとわかってしまう。正答率がかなり高くなる問題になってしまう。 • 各種のラインで水を入れるような管理や毎日の交換を必要とするものが思いつかないため。	

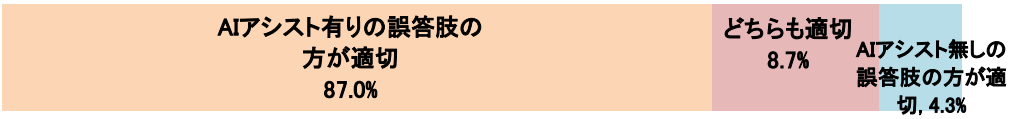
問題 5	以下は世界保健機関〈WHO〉の定義における健康の概念の憲章前文の一部である。「健康とは、病気ではないとか、弱っていないということではなく、肉体的にも、精神的にも、そして〔　〕にも、すべてが満たされた状態にあることをいう」〔　〕に当てはまる語句はどれか。	正答選択肢：社会的
誤答肢セット 1	AI アシスト有り 誤答肢 ◆：AI 作成誤答肢	経済的◆ 環境的◆ 文化的◆
誤答肢セット 2	AI アシスト無し 誤答肢	心理的 経済的 年齢的
誤答肢セットの適切性評価		
誤答肢評価の理由		
誤答肢セット1(AI アシスト有)が適切	• 誤答肢 2 の方が間違いがわかる • 誤答肢 1 の方が、社会的に近い • 精神と心理的は重複 • 年齢的という選択肢があきらかにふさわしくないので • 誤答肢 1 の方が迷いやすい回答肢が多いと思う。2の年齢的は誤答であると容易に分かると思う。 • 社会的という正答に対して、文化的と誤答しやすいのでないかと推測するため。 • 誤答肢 2 の年齢的という意味が適正ではないと感じた • 3 つのキーワードから（ ）に入るとイメージしそうだから • 誤答肢 2 は簡単すぎる • 誤答肢 2 内の選択肢に年齢的があることで、誤答肢 1 より魅惑肢と感じられないから。 • 誤答肢 1 の方が抽象度が統一されているため。	
誤答肢セット2(AI アシスト無)が適切	• 満たされるという文章から、心理的という言葉で迷いそうと考えたため。語として不適切であり、カテーテルは 30cm 程度何の目安になるのかがわからない。水平仰臥位という表現も聞きなれない。	
どちらも適切	• この問題については、すべての学生は知っていると思うため間違えないだろう • 不適切な回答ないため • どちらも知識がないと解けない • WHO 憲章前文の内容と類似しているものもあるが、知らなければ選べないと考えたから。 • どちらの誤答肢セットも、文脈上正解の可能性がありそうだと感じたため、どちらもよいかと思いました。 • 前文の文言の答えは決まっているため、社会的に似た内容であれば基本的には良いと思う。	
どちらも不適切	• 環境的、年齢的という用語があるのか不明。経済的、文化的、心理的が適当ではないか。	

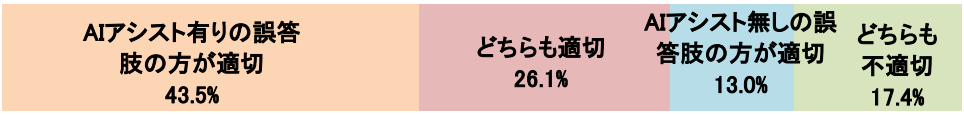
問題 6	令和元年（ 2019 年）の国民生活基礎調査における通院者率が男女ともに最も高いのはどれか。		正答選択肢：高血圧症
誤答肢セット 1	AI アシスト有り 誤答肢 ◆：AI 作成誤答肢	糖尿病◆ 心疾患◆ 脳血管疾患◆	
誤答肢セット 2	AI アシスト無し 誤答肢	糖尿病 歯の病気 脂質異常症	
誤答肢セットの適切性評価			
誤答肢評価の理由			
誤答肢セット1（AI アシスト有）が適切	• 疾患名でよいから • 死因と混同して間違える学生はいる • 糖尿病と誤答しそうだが、死因と混同して心疾患・脳血管疾患と答える学生もいそう。 • 既往歴として含めた考えを持つ可能性がある • 疾患名が明確に示されている • 誤答肢 2 は問題文をきちんと読まない学生の正解率が上がると思ったからです。勉強をしている学生にとっては誤答肢 2 の方が難易度は高いと思いますが、勉強をしていない学生の方が正解率が上がるというのは本来の趣旨に沿っていないのではないかと思いました。 • 誤答肢 1 は通院が必要な疾患が挙がっているが、誤答肢 2 は歯の病気の範囲が広く不明、歯科への通院は高血圧より多くはないのか？		
誤答肢セット2（AI アシスト無）が適切	• 歯の病気、脂質異常は、正確なデータが無いと併せ持った病気で迷うと思われる。 • 糖尿病は魅惑詩肢となりうる、歯の病気も一般的であり両者が含まれるため • 疾患を並べるより症候群も入れたほうが良いと感じた • 大きな疾患というよりも、糖尿病患者や生活習慣からの症状を考えて答えるのではないかと考えた。 • 上位傷病に含まれる選択肢であるため • 心疾患や脳血管疾患の選択肢は、診断がついている人が少ないと思うので誤答肢1は難易度が低いとおもったため。		
どちらも適切	• 簡易		
どちらも不適切	（回答なし）		

問題 7	介護保険法に基づく地域支援事業である包括的支援事業の中核的機関はどれか。	正答選択肢：地域包括支援センター
誤答肢セット 1	AI アシスト有り 誤答肢 ◆：AI 作成誤答肢	保健所◆ 介護保険事業所◆ 介護老人保健施設◆
誤答肢セット 2	AI アシスト無し 誤答肢	福祉事務所 社会福祉協議会 社会保障審議会
誤答肢セットの適切性評価		
誤答肢評価の理由		
誤答肢セット1（AI アシスト有）が適切	• 誤答肢 2 は、迷わなさそうです • 介護というワードが多いため迷うと考える。模試などで復習をする時は一つ一つ事業所の特徴について抑え直しができそうである。 • 事務所、協議会、審議会が地域包括支援センターとの比較としてふさわしくない。 • 誤答肢1の方が「介護」という言葉に影響され迷うように思う。 • 誤答肢 2 の方が聞き馴染みのない選択肢ではないかと感じ、正答がわからないと消去法では回答しづらいと考えたため。 • 介護に関する施設の方が迷いが生じ、正しい回答を持つことが必要なため、その選択肢が良いと感じた • 違いが明確 • 誤答肢 1 の魅惑肢に事業所とあることで、問題文にある事業という言葉から、正答として選択されやすいのではないかと思ったため。 • 問題文に介護保険法に基づく、という文言があるため、誤答肢 2 の福祉よりセット誤答肢 1 の方が魅惑肢になると思うため。 • 近いカテゴリに含まれる言葉のため、誤答肢 1 の方が良いと思った。	
誤答肢セット2（AI アシスト無）が適切	• 学生に、地域包括支援センターと社会福祉協議会の区別をつけて欲しい • 知識がないとそれぞれの機関の役割が解らないから。	
どちらも適切	• 法的根拠のものなので • 不適切な回答ないため • 問題文にある「中核的機関」を問われると、正しい知識がなければ正しい答えを選ぶことはできないと考えたから。 • 問題文の「介護保険法に基づく」という文章から、介護にかかわる機関であることは読み取れますので、「介護」と入っている誤答肢1は知識がないと間違えやすい所だと思いました。誤答肢 2 は、事業の内容が分からないと答えられないため良いと思いました。 • どちらも魅惑肢になりうる	
どちらも不適切	• セットの内容からは選択ができないのではないかと考えたため。	

問題 8	要介護状態の区分の審査判定業務を行うのはどれか。	正答選択肢：介護認定審査会
誤答肢セット1	AI アシスト有り誤答肢 ◆：AI 作成誤答肢	介護保険審査会◆ 診療報酬審査会◆ 保健所◆
誤答肢セット2	AI アシスト無し誤答肢	介護保険審査会 老人福祉センター 市町村保健センター
誤答肢セットの適切性評価		
誤答肢評価の理由		
誤答肢セット1(AI アシスト有)が適切	• 認定審査会は、保健センターは、あり得ないので、外せばいいかと思っています • 審査という文言のつくものが多いのでそちらを選択したくなるように感じたため。	
誤答肢セット2(AI アシスト無)が適切	• 分かっていないと市町村の中のどこの？の迷うかもしれない。 • センターが惑わされやすい • 介護認定のことを学習した時に、介護認定がおりて介護保険を使う、申請するときに役所に届け出るなどのイメージがあるのではないかと考えるから • 知識のない学生にとっては、選択肢全てが確からしく思えるかもしれず、消去法も難しそうだから。 • 誤答肢 1 の選択肢のうち 3 つが「～審査会」です。つまり、保健所は明らかに誤答肢です。あとは、「診療報酬」でないことはわかるので、事実上 2 択問題になってしまうため。	
どちらも適切	• 介護保健と名称があるので正しい知識がないと判断ができない • 曖昧な内容が混在しているため • 両方とも関係しそうな機関が挙げられている	
どちらも不適切	• 保健所と保健センターの対策はかなりするので、間違えることは少ない • 誤答肢 1 の選択肢の 1 番最後を市町村保健センターにすると良いと感じた • 介護保険審査は誤解しやすい	

問題 9	利尿薬の使用目的で正しいのはどれか。	正答選択肢：血圧を下げる。
誤答肢セット1	AI アシスト有り誤答肢 ◆：AI 作成誤答肢	体温を下げる。◆ 血糖値を下げる。◆ コレステロールを下げる。◆
誤答肢セット2	AI アシスト無し誤答肢	血糖を下げる。 血液を凝固する。 免疫を抑制する。
誤答肢セットの適切性評価		
誤答肢評価の理由		
誤答肢セット1(AI アシスト有)が適切	• 免疫抑制は曖昧 • 下げるで統一してるから • 誤答肢 2 にくらべ誤答肢 1 のほうが、回答者に迷いが生じると考えたから。正しい知識がなければ答えることは難しいと感じた。 • 誤答肢 2 は血糖降下薬など、言葉だけで回答できそう。 • 排尿が促進されることで、体内でどのような反応があるのかを考えてもらう選択肢であればよいと思いましたので、どちらの選択肢でも良いのではないかと考えました。しかし、誤答肢 2 は血糖降下薬、血液凝固薬、免疫抑制剤など他の薬の名前を連想しやすく、知識がなくても答えられてしまうかもしれないと思い、誤答肢 1 にしました。 • 誤答肢 1 の誤答肢すべてに下げるという言葉が使用されており、回答に迷いやすいと感じたため。 • 誤答肢 2 は簡単すぎ • 回答が下げるのため、下げるで統一したほうが考えると思った。知識がなく体から何か出ることで下げるイメージであればコレステロールと迷うだろうし、血圧・血糖の似た言葉でも迷うと思う。	
誤答肢セット2(AI アシスト無)が適切	• 利尿薬であるため尿排泄を促すことは容易に判断できると思うが、そのため誤答肢2にある「血液を凝固する」と誤答しやすいように思う。 • 誤答肢1は、排尿により体温が下がることがあるのではないかと思う。誤答肢 2 は血糖を下げると迷いそうという選択で良いと思う • 体水分量が減少することで起こることが選択肢になっているので、魅惑肢として適当	
どちらも適切	• 不適切な回答ないため • 誤答肢 1 の方が文末の統一感があるが、難易度としてはどちらも変わらないと思う。	
どちらも不適切	• 利尿薬に関係ない選択肢 • 迷わせるような選択肢ではないと考える。サービス問題として出題するのであればどちらも適当であると考える。 • わかりやすい問題でもあるから。 • 関連性がなさすぎ • セットの内容からは選択ができないのではないかと考えたため。 • 利尿薬は学生でも実習でよく聞く薬剤であるため理解している回答者が多いと思われ、どちらのセットも魅惑肢にならないと思った。	

問題 10	ワルファリンカリウムの作用に影響を及ぼす食品はどれか。	正答選択肢：納豆
誤答肢セット1	AI アシスト有り誤答肢 ◆：AI 作成誤答肢	牛乳◆ バナナ◆ グレープフルーツ
誤答肢セット2	AI アシスト無し誤答肢	チーズ 米飯 牛乳
誤答肢セットの適切性評価		
誤答肢評価の理由		
誤答肢セット1(AI アシスト有)が適切	• 内服薬の制限でグレープフルーツがよく出るので間違いやすい。 • このままで良いと思う • グレープフルーツと混同する学生あり • 牛乳やグレープフルーツなど講義で取り扱いがあるものなので正しく理解していないと選択できないため • グレープルーツが入ったほうがいいと感じた • 何となく膜を作ってしまいそうなものと、グレープフルーツというワードで選択する • 薬理学で降圧薬とグレープフルーツのことを学習しており混同しそう。バナナはカリウムが多いと知っていると、ミスリードされそう。 • グレープフルーツやバナナを選ぶ学生がいるのではないか。 • カルシウム拮抗剤との関連でグレープフルーツと迷う可能性がある • グレープフルーツが知識がないと誤って解答しやすいと感じたため • グレープフルーツが入るため • 勉強をしている学生は、グレープフルーツと納豆は覚えているため。誤答肢 2 は牛乳とチーズはほぼ同意、米飯はなぜ魅惑肢になるのか、私自身が理解できないですし、学生も選択しないと思います。 • 牛乳とチーズは両方とも乳製品であり、米飯は可能性が低くすぎるので選ぶことはほぼなさそう。 • アレルギーなどと混同しやすい選択肢のため。	
誤答肢セット2(AI アシスト無)が適切	• 影響を及ぼす食品の内容が明らかに異なるため	
どちらも適切	• 特になし • 簡易	
どちらも不適切	(回答なし)	

問題 11	ストレス下における生体反応で正しいのはどれか。	正答選択肢：血圧上昇
誤答肢セット 1	AI アシスト有り 誤答肢 ◆：AI 作成誤答肢	血糖値低下◆ 呼吸数減少◆ 瞳孔散大◆
誤答肢セット 2	AI アシスト無し 誤答肢	尿量増加 体温低下 血糖低下
誤答肢セットの適切性評価		
誤答肢評価の理由		
誤答肢セット1(AI アシスト有)が適切	• 迷いそう • 呼吸について入っているので交感神経系を連想させるので迷う学生もいそうだと考えるため。 • 交感神経、副交感神経のはたらきを理解してないと回答できないため • 「血糖値低下」または「低血糖」はよく聞くのですが、「血糖低下」はあまり聞いたことがなかったため誤答肢1が良いと思いました。 • 誤答肢 1 の方が回答に迷いやすいと感じたため。 • 瞳孔散大ぐらいしかひっかからない • ストレス下での反応にないことが明確でわかりやすい回答。	
誤答肢セット2(AI アシスト無)が適切	• 自分の体でイメージしやすい • 誤答肢 2 の体温低下や尿量増加は迷いやすい回答肢だと思う。	
どちらも適切	• 不適切な回答ないため • どちらも知識がないと解けない • 「ストレス」という言葉を見たとき、多くの回答者は交感神経を思い浮かべると思う。どの誤答肢も交感神経に関連がある内容と考えるため、正しい知識がなければ答えることができないと感じる。 • どちらも同じような内容でどちらでも良いと思う • どちらも問題ない • 自律神経の作用の選択肢としてどちらも妥当だと思う。	
どちらも不適切	• どちらも答えがわかりやすい • セットの内容にある減少とストレスは繋がらないため。 • 心拍数や食欲に関する魅惑肢があると、正しい知識の習得状況がわかるのではないかと思ったため。	

問題 12	訪問看護ステーションの管理者になることができるのはどれか。	正答選択肢：看護師
誤答肢セット1	AI アシスト有り誤答肢 ◆：AI 作成誤答肢	理学療法士◆ 薬剤師◆ ケアマネージャー
誤答肢セット2	AI アシスト無し誤答肢	医師 市町村長 都道府県知事
誤答肢セットの適切性評価		
誤答肢評価の理由		
誤答肢セット1(AI アシスト有)が適切	• 職種などで、した方がいいから • 理学療法士やケアマネと混同する学生はいる • 誤答肢 2 は違うことが明確すぎる • 選択肢が全て医療従事者であるほうが選択しにくい。また、似たような過去問があった。 • ケアマネジャーを選択する学生が一定数いそう。 • 紛らわしさとしては選択肢 1 の方が魅惑であると感じた • 訪問看護ステーションがどのようなものかを理解していれば、市町村長や都道府県知事を選ぶことはないと思います。ですが、誤答肢 1 に「医師」がある方が病院長が医師なので、魅惑肢となるのではないかと思いました。 • 村長や知事は同類であり、設問との関係が薄く魅惑肢にはなりにくい	
誤答肢セット2(AI アシスト無)が適切	• 管理者なので看護師の上の医師や組織が必要と間違いやすい。 • 医師はないと誰しも考える • 誤答肢 1 であれば誤っているものとして除外しやすい　医師はできそう　市町村も公立の機関もありそうなので学生は間違えそう • 管理の問題のため、項目が適切なため • 医師なら良いと考えそう • 訪問看護ステーションなので、理学療法士や薬剤師はなれないと思いやすいので。	
どちらも適切	• セット1：ケアマネージャー、セット2：医師、と考えそう。管理者というイメージから。 • シャッフルしやすい	
どちらも不適切	（理由なし）	

問題 13	地域包括支援センターの機能はどれか。	正答選択肢：介護予防ケアマネジメント
誤答肢セット1	AI アシスト有り誤答肢 ◆：AI 作成誤答肢	介護認定審査◆ 医療設備の提供◆ 高齢者福祉サービスの提供◆
誤答肢セット2	AI アシスト無し誤答肢	配食サービス 金銭の管理 訪問介護の実施
誤答肢セットの適切性評価		
誤答肢評価の理由		
誤答肢セット1(AI アシスト有)が適切	• 誤答肢 2 は明らかに役割ではないため選択肢にならないと思うため • 正しい知識がないと、間違いやすい項目だから • 誤答肢 2 は実際のサービスの内容を想起させるような内容であると感じるため、誤答肢 1 のほうが知識がなければ解くことはできないと考える • 1の介護認定審査は迷いやすい回答肢だと思う。 • 誤答肢 1 の方が迷いやすい選択肢が多いと感じたため。また、誤答肢 2 は正答選択肢に対して、他の魅惑肢が直接的に提供される具体的な内容のため、正答選択肢が目立っていると感じたため。 • 正しい知識を問われていると感じる • 誤答肢2は実際のケアであるため、魅惑肢にはなりえないのではないか • 誤答肢 1 の介護認定審査など、それらしいような魅惑肢があると感じたため？ • 問われているのが機能なのに対し、誤答肢 1 の方が抽象度が揃っており、誤答肢 2 は魅惑肢が具体的すぎるため。 • 言葉の抽象度が誤答肢 1 の方があっていると思った	
誤答肢セット2(AI アシスト無)が適切	• 高齢者福祉サービスは曖昧 • 誤答肢 2 は、正答よりも抽象度が低いため知識がなくても解けそうだと思いました。	
どちらも適切	• 学生の知識があやふやだと、どれも選択しそう • 不適切な回答ないため • どちらも知識がないと、選択肢を 2 択にしぼる事ができない。 • 強いていうなら 1 の方が難しいから 1 だが、どれもサービスとして存在していて耳馴染みがある • 介護認定や、訪問介護で迷うと思う	
どちらも不適切	• 特に誤答肢 1 医療設備の提供、誤答肢 2 金銭の管理は選択はしないと考える。	

問題 14	非言語的コミュニケーションはどれか	正答選択肢：ジェスチャー
誤答肢セット1	AI アシスト有り 誤答肢 ◆：AI 作成誤答肢	手紙◆ 電話 点字
誤答肢セット2	AI アシスト無し 誤答肢	点字 手話 翻訳機の使用
誤答肢セットの適切性評価		
誤答肢評価の理由		
誤答肢セット1（AI アシスト有）が適切	• 手紙（やり取りの場に双方がいなくても成立する）と点字（言語のイメージが湧かないかも）に惑わされそう。	
誤答肢セット2（AI アシスト無）が適切	• 発語を繰り返さないグループでうっかり間違いがある。 • 手話を混同する学生はいる • 手話も非言語的コミュニケーションの一つと考えるため • 声を使う発語という意味で認識した場合に迷うかもしれない • 翻訳機の認識がどうか • 手紙、電話は言葉を使うのでセット2の方が紛らわしいと感じたため。	
どちらも適切	（理由なし）	
どちらも不適切	• 非コミュニケーションの問いに合うようにしてほしい • わかりにくいと感じた • 看護の領域を超えていると思われるため • ジェスチャーだけが浮いている印象を受けます	

問題 15	病室の環境で適切なのはどれか。	正答選択肢：冬季の室温は 20～22℃とする。
誤答肢セット 1	AI アシスト有り 誤答肢 ◆：AI 作成誤答肢	夏季の室温は 30～32℃とする。◆ 照度は 1000 ルクス以上である。◆ 湿度は 75～80%とする。
誤答肢セット 2	AI アシスト無し 誤答肢	夜間の騒音は 100dB 以下とする。 昼間の病室の推奨照度は 1,000 ルクスとする。 病室の床面積は患者 1 人につき 3.0 ㎡以上とする。
誤答肢セットの適切性評価		
誤答肢評価の理由		
誤答肢セット1(AI アシスト有)が適切	• 特に誤答肢 1 湿度の数値は迷う選択肢にはならない。	
誤答肢セット2(AI アシスト無)が適切	• 一（夜間の騒音）は、明らかに違う選択肢 • セット 1 は難易度が簡単なため • セット 1 は湿度の数値、室温などが明らかに違うため • 療養環境に必要な項目が揃っている • 文章が長くセット 1 の方が解きやすい • 1 は悩む余地がなくオーバー過ぎる • 誤答肢セット1は環境の基準値を知っていれば解ける問題ですが、誤答肢2は、昼間と夜間での環境を聞いていたり、床面積も聞いていて幅広い知識が求められるため誤答肢 2 のほうが良いと思いました。 • セット 1 は専門知識がなくとも常識の範囲内で消去法で回答できると考えるため。 • 正答と同じような文章構成であるため • セット1は明らかに誤答が多く、魅惑肢になりえない • セット 1 に夏季の室温の選択肢があることで、魅惑肢としての精度が下がるように感じたため。	
どちらも適切	• 問題 1 は照度と迷いそう。問題 2 は知識がないと答えられない • 誤答肢 1、2 ともに、基礎的知識だがわかっていないと解けない問題であると考えたため。 • 1の湿度や室温は明らかに誤答だと分かりやすい。また照度 1000 ルクス以上の「以上」が付いているため、これも明らかに誤答と分かりやすい。 • どちらも病室環境について妥当な選択肢であると思うため。	
どちらも不適切	• 数値が大幅 • セット 1 は照度のみ～であるとなっており、気になった。セット 2 は夜間の騒音はという言葉があいまいでよくないと思った。夜間でも就寝時かそうでないかどう想像するかが異なると思う。	

問題 16	薬物が肝臓を通過した後に標的器官に作用する与薬法はどれか。	正答選択肢:経口与薬
誤答肢セット1	AI アシスト有り誤答肢 ◆:AI 作成誤答肢	静脈内注射◆ 筋肉内注射◆ 経皮与薬◆
誤答肢セット2	AI アシスト無し誤答肢	皮下注射 口腔内与薬 直腸内与薬
誤答肢セットの適切性評価		
誤答肢評価の理由		
誤答肢セット1(AI アシスト有)が適切	• 静脈から門脈に直に運ばれ、それから作用するイメージを持っているかもしれない。 • 消化管のルートからすると肝臓よりも直腸のほうが下流のイメージがあるため。	
誤答肢セット2(AI アシスト無)が適切	• 経口与薬を経口内与薬の違いが不明瞭で間違う可能性あり。 • 内服を問うため、こちらで良い • 口腔内投与と混同する学生はいそう • 注射は選択肢に不要 • 肝臓と門脈の理解をしている場合、腸管からの吸収を考える可能性がある • 口腔という表現で悩む学生がいそう • 口腔内が吸収経路を誤って認識していると紛らわしい選択肢になると考えるため • 誤答肢 1 は注射が誤答肢 2 つ入っているため	
どちらも適切	• どちらも正しい知識がないと答えることができない • 選択内容が適切なため	
どちらも不適切	• 知識が必要 • 「わからない」という選択肢がないので「どちらも不適切」にしました。「注射」という概念と「与薬」という概念が並列可能なのかご検討ください。	

表 28　マルチメディアを活用した問題に対する感想（学生）

カテゴリ	サブカテゴリ	具体的例
文章問題との比較	状況は文章よりも理解しやすい	状況設定問題が文ではなくて動画の方がイメージしやすく、やりやすかった。
		字面で読むよりも何を問われているかが動画のほうが明確、伝わりやすかった。
	必要な情報を得るのは文章問題よりも時間がかかる	問題の内容によっては、文字だと大事なところだけを先読みしてポイントをつかんで速く読んで答えられるのが、動画だと最初から最後まできっちり見ないと答えが得られないので、時間が心配になる。
		問題の難易度はそんなに難しくはないが、今までは紙だったら画像が 4 枚あってそこから選ぶから、ぱっと見てぱっと動画だったので、「おっと」って感じで巻き戻したりとかしてだったので、ちょっと時間はかかった。
試験対策，学習に関する影響	臨床に生かせる形で学習できる	国試の勉強を文字でしていると、国試にはいいが臨床にはいかせないので、映像で見た方が実習の経験も役に立つので映像もあった方がいい。
		映像の方が実践ではいかせるが、人に伝える時に言葉で表さなければいかないので、紙で単語を覚える工程が大事だと思う。
		自分が働く現場を想定したら、コンピュータのほうが実際の現場に近い状態での音も聞けるし、そういうので分かりやすさでいったらコンピュータのほうがいいかなって思った。

表 29　回答形式についての感想（学生）

カテゴリ	サブカテゴリ	具体例
回答形式の違いに関して	転記が不要なので解答に集中できる	紙問題では解答を照らし合わせる時にミスしてないか不安になるが、CBT 試験の方がミスは防げると思う。
		紙だと間違えて色塗ったりとか、塗ってなかったりとかがあると思う。コンピュータだと、記入漏れとか間違えて選ぶとかは、なくなっていいのかなと思った。
	クリックで問題に飛べるので楽	スキップした問題を一覧で見て戻ることができて簡単だった。
	メモ機能の必要性	問題に印をつけられる機能があればいいが、細かい機能が増えると使いこなせない人とたちが出てくるであろうから難しいと思った。
		紙のテストは印やメモを書きこむことができる点がよいが、CBT でもブックマークやメモの機能、戻れる機能があるか、手元にメモできる紙があればよい。
		滴下計算とかあったら、ちょっと難しいかな。筆算したい。計算問題ができない。
	時間配分	時間制限ある中で動画の問題が続くと、（進みが遅く感じて）早くしてよっていう感じになる。
		紙問題であれば最初に問題全体を見て計算問題はどこにあるか等を確認できるが、CBT ではどこに動画問題があるかわからないから時間配分がわかりにくい。
	試験形式の説明	動画は 1 回しか見られないと思った。リスニング試験のように仕組み（2 回読むなど）の説明がないと不安だ。試験をやりながらこれできるのだとわかってきた感じだった。
		呼吸音の問題：聞き取りにくかったが、PC 自体の音量が小さかったのだと思う。英語のリスニング試験のように、テスト前に設定の方法を伝えたり、チュートリアルとしてお試し問題や動画があったりした方が安心だ。
視認性について	紙の方が読みやすい	紙の方が慣れているため文が読みやすい。
	画面を見ることによる疲労	一定時間パソコンをずっと見るのがつらいので、紙の方が解きやすい。
		長時間になると目がチカチカしたり、頭や背中が痛くなったりすると思う。
		状況設定問題の長文がいっぱいある問題とかだったら、一遍に集中するから目疲れるかなっていうのはちょっと感じた。
マルチメディアについて	動画・音声の視聴形式	動画を倍速で進めることができたら助かる。
		動画の解答の選択肢が 1 から 4 まで 1 本の動画なので、1 つだけ見返したい時には全部見るか、シークバーをこの辺りかと検討をつけて動かさないといけない。選択肢の動画が分かれていたらもっと見やすいと思う。
		動画は最低限 2 回は再生できるようにしてほしい。再生回数の上限がなければありがたい。
	イヤホンについて	イヤホンを長時間つけると気持ち悪くなるので考えていただきたい。
		イヤホンは耳に合うものと合わないものがある。自分は柔らかいタイプが合わないので、選べるといい。

表 30　CBT 形式になった場合の学習・試験対策に有用・必要だと思うもの（学生）

カテゴリ	サブカテゴリ	具体例
現状行われているもの	シミュレータ・マルチメディア教材による学習	急性期の実習時にシミュレータの人形で脈を測ったりモニターを取ったりしたこと。
		基礎のフィジカルアセスメントのホームページで音を聞いたこと。
	演習・実習	基礎看護の授業の時に見たり、自己学習したりした丸善の EVO の動画に似ていた。
		手袋は基礎看護で、BLS は授業で、小児の授業では実際に赤ちゃんと大人に練習した。
		何度も練習した手技は流れを覚えているので、このような試験の問題を解くスピードに関わってくると思う。
追加で必要だと思うもの		心電図のモニターを触って実際に動く波形を読むこと。
		今回わからなかった問題もマネキンや自分の体を使いながら教えてもらったら、答えられる問題が多いと思う。
		心臓の音とかって、やっぱり授業で何回か聞いたぐらいしか印象がなかったから、腸蠕動とかもそうだが、あの辺は結構授業の中で繰り返し聴き慣れとかないと、しんどいところがあるかなと思った。

表 31　CBT 化の利点・欠点についての意見のまとめ（教員）

カテゴリ	サブカテゴリ	具体的なコメント
CBT 化のメリット	マルチメディア問題が使用できる	紙で書いてあるものは、情報がすごい限られてる（見えない、聞こえない）が、雰囲気や表情も含めたビデオ教材から設問にすることは、学生さんたちはそれもアセスメントすることができる。文章で書くことは、それを、もうすでに出題者が書いてしまっているか書かないかというところが出てきてしまうから、より学生さんの能力をアセスメントする意味では、情報が豊かな分だけ、良いアセスメントが試験の目的として、できるのではないかと思う。
		いわゆる文章だけで知識を問うのは非常に難しくなってきていることと、あとは学生の、どこまで理解しているかは、動画、例えばサンプルの滅菌手袋の問題があったが、あれは非常にいいなってと思った。どうしてもやっぱ画像だけだと細かいところ伝え切れないが、動画だと「あ、ここが違うな」っていうのが、ほんとに知識が試される問題なので非常にいいなと思った。
	複数回実施が可能になる	一番のメリットは、年 1 回ではなくなることが一番のメリットと思う。コロナの時に、もしかしたら試験自体ができないかもしれないと思った時に、何回かに分けてできるとか。ただ、そうなった時に、結局、CAT ではないんですよね。CBT。 計算はしないで、難易度は計算するにしても、そうなった時に、じゃあ、複数やるなら、問題セットの難易度をどう調整しておくのかとか、そういうことは考えなきゃいけないけれど。
		将来的には、やっぱりアメリカみたいに何個か割と試験会場を多くして、そこの中で年に何回かやれるみたいなことをやるほうが、人の確保とか、1 回の受験で落ちちゃった人の再チャレンジとかにも効果的かなと思うので、コンピュータを利用した試験になることには賛成をしている。
		あとは多分これが稼働していくと、一回免許取った人たちの更新にもつながっていくのかなって思って。臨床の数値ってほんとに忙しい中でいろんな学習会開いたりとかしているが、基本的な知識に立ち戻ることもまた大事かなと思って。そういう意味では、更新制度っていうのも日本でどの程度進むか分からないが、そういうものにつながっていけば、看護師全体としての自己研鑽とかレベルが上がっていくことにつながっていくのかなと思っているので。
	採点等の利便性	基本的には紙ベースじゃないほうがいいというのは、その集計とか採点とかの利便性もあります
CBT 化のデメリット	マルチメディア問題による情報量の増大の試験への影響	一方で、映像みたいなものから、たくさんの情報が与えられてくるわけで、それをどう捉えるかということが、均一・平等にできるのかというところに関しては、難しいのかなと思う。
	作問負担の増大	作問する側としては大変と思う。適切で答えがばらつかないような動画を作れるんだろうかとか、一年一年で問題を作るわけですけど、その期間の中で動画を作らなきゃいけないのは、ほんと大変だなと思う。
		問題作りも大量にいいのを作らないと。全国の教員全員に 1 人 10 問とか、20 問とか、何でもいいから作って出せって。マストでやってくださいぐらいじゃないと無理かもしれない。
	公平性の担保が難しい	看護師はやっぱり医師よりも N が多いということと、やはり看護基礎教育ですごく幅がいろんな大学から専門学校からいろんなルートで看護師になるところから考えると、看護基礎教育の期間もやっぱりインフラ・設備とかがさまざまある中で、一定の全ての受験生の公平性を保った準備を、インフラの準備ができるかなという点でちょっと疑念があって。
	インフラの整備が難しい	看護師はやっぱり医師よりも N が多いということと、やはり看護基礎教育ですごく幅がいろんな大学から専門学校からいろんなルートで看護師になるところから考えると、看護基礎教育の期間もやっぱりインフラ・設備とかがさまざまある中で、一定の全ての受験生の公平性を保った準備を、インフラの準備ができるかなという点でちょっと疑念があって。

		何か現実的に各一人受験生に PC をそろえてするっていうのが現実的に何ていうか、予算的なところであるとか、そういうのって設定は可能なのかはちょっと疑問はある。
		指定された会場に集合して備え付けのディスプレイうんぬんというところでは、相当数の機材が必要になると思われて、その設置にかかる準備とか、配線とか、会場の配置とか確保、その辺りはどうなんだろうという、すごく懸念されると思った。
	実施費用が高額になるのではないか	受験料金がこれをやるとしたら、お金がかかってきた時に、どのくらいになるのかなっていうのは、ちょっと懸念する。
	漏洩等のセキュリティ	欠点はシステム上の問題で、どこまでいっても紙でも一緒だが、やっぱりセキュリティの問題とか、デジタルデータになるので一度流出すると分からない部分があること。一般的なデジタルのリスクは、ちょっとどれだけ仕組みでカバーできるかっていうところは課題かと思った。
	紙でできることの再現に関する問題	今現在ペーパーで試験を受けている学生が、思考の整理をしながら線を引いたりとかして自分がどういう解答を導き出すかというのをマークシート以外に問題用紙に書いているので、そういう作業をパソコン上でできないのでメモ用紙とか持ち込みが可なのか、そういう問題をただ単にチェックするというのに慣れてない学生なので、現状は難しいかなと。

表 32　マルチメディアを利用した問題に対する評価（教員）

カテゴリ	サブカテゴリ	具体的なコメント
マルチメディアを利用した問題について	難易度に対する評価	現在の国家試験の難易度と比較して難しくなると思う。あの波形は見せてないですよね、心電図の。だから傾向と対策がちょっと変わるかも。音はない。
		難易度的には、そんなにものすごい高いっていうわけでもないと思って。逆に動画になったりとかしてるので、正確な形で問題としても作れますし、というのはとても感じた。
		動画に関してはやっぱり先ほども言ったように視点をどこに置いてるかというところで変わってくるかなと思うので、難易度はやっぱり高いのかなとは。看護の経験というか、そういった場面というものをやっぱり学生が知らないというか、映像とかそういったもので学習はしていくと思うんですけど、そういった経験というのがやっぱりないので難しいのかなっていうのは思います。
		動画とか音声とか、あと普通の 4 択の問題があったと思うが、難易度自体は簡単だったかなと思っているが、動画のほうは先ほども話があったが、着目するところがいっぱいあって、一体自分は何を見なきゃいけないのか、もう一回見てみようという、その時間がかかるので、解答に時間がかかって、もしかしたら終わるまでの時間がかかる、足りないことにもなるんじゃないかなと思った。
	能力を評価する試験としての適切さ	多角的に判断するのは、今まで紙、文章で出せない問題をちゃんと出せてて、いいと思った。国家試験としては今よりふさわしい。
		より実践に強い人が合格しやすくなると思って、すごく面白いと思いました。
		今の国家試験問題を踏まえて作成された動画を使った問題としては、すごく適してると思う。
		やはり状況判断とか、それから視覚的な情報から得たものでの作問というのは、やはり文章では出せなかった問題だと思いますので、臨床実践により近いといいますか、実践内容を問うこともできるので、紙のベースよりも問う能力は広い範囲で問えるのではないかと思いました。
	解答時間・問題数	今みたいに 1 日やってたら大変かも。 今現在も、国家試験免許を与えるにはこれだけの知識量を取るという、そこは割と検証されて、あの時間の作問になったのか。全然関わったことがないので、やっぱり領域ごとに何問っていう設定で、それが結局ああなってるのかなとは思うが、そういう最適な時間というか、最も経済的な時間とか問題数とかっていうのが、コンピュータ化していくことによって進むのかなとか思うとなるといいなと、でも、確かに、ずっと見るのは大変かもしれない。
	動画の再生回数	動画に関しては流れていってしまうので、あれはもう一回見られるようになっているのか。何度でも見られるのであればいいかと思った。流れていくと、自分が注目すべき場所がどこなのかっていうのを、一瞬分かんなくなる人もいるから。
	視覚・聴覚の多様性について	色覚多様性の人って、カラー問題がいくつかありましたよね、確か。あれ、どうしてるんだろうなと思う。マルチメディアが使えるようになると、音も聞こえにくい人の問題が出てくるから、音と色が見えない人へ、見え方が違う人への配慮をどうするかっていうのを考えなきゃいけないなと思う。
	外国人受験者への対応について	目が見えないうんぬんでやると、外国の人が確かに難しいだろうなと、会話形式で出題した時なんかは、字幕付けるのか、でもこれぐらいは分かる人でないといけないってやるレベルの会話だけにするのかという、そのあたりやっぱり合格基準の設定が改めて難しいと思った。 全部、平仮名振る感じなので、上に字幕を付けるっていう配慮をするのかしないのかというあたり。

マルチメディア問題作成時の課題	モデルのプライバシー	子どもだと倫理的なところとかもありますし、今のテストだと目とかも全部写っている。通常消したりしてるはずなので、そのあたりどういうふうにするのかなって、SPさん、模擬患者さんのプロフェッショナルみたいな人がやってくれると上手かもしれない。
	誤答肢の動画等を作成する労力がかかる	今、テキストには該当の動画いくらでもあるんだけど、あと 3 問、間違った場面を作らなきゃいけないのはすごく大変。だから動画があるから問題作りやすいんじゃないの？というのはちょっと違うと思う。素材はある。正しい素材はあるけど。誤った素材を準備するのが大かな。
	動画等の解釈の多様性に伴う問題	状況設定問題、動画でやろうかと思うと非常に難しいだろうなとか思って。細部まで気をつけないといけないから、危険だろうなとか思って。
		紙で文章で作った問題よりも、動画のほうがいろんなふうに読めるから、より答えの幅が広くなって、終わった後に、疑義がつきやすくなる懸念がある。そんなことが起こらないように、事前にしっかりと動画で求めることとか、こういう問題の場合は、こういうことを意図してるっていうことを、しっかりと受験生も、看護教育機関も、準備できるような情報提供が必要

表 33　問題プール作成のための過去問題非公開化に対する意見（教員）

カテゴリ	具体的なコメント
学生の不安	過去問が非公開になることに対する学生さんの不安とかは、きっとあると思う。
予備校等による対策	結局は公開しないとなったら、多分学生が分担して記憶してみんなで持ち帰って、大学なり学校が教えるっていうのは、やると思う。
	受ける人たちのための国家試験対策の観点としては、恐らく業者が頑張って、お金かけてくると思うので、その辺でカバーはできるのかなと思う。
国際的には非公開が一般的	テスト学会系の先生たちとずっと話をしてて、そもそも日本が公開してることのほうがおかしいんだと、ずっと聞いてたので非公開で問題ないと思う。
過去問対策も学習の一部である	落とすための試験じゃなくて勉強させるための試験だからやはり過去問から何度も解いて復習して知識を蓄えることを、今は国家試験受験するために受験生はみんなやってるのを非公開にしたらそれはできなくなるっていうことはやはりちょっと本来の趣旨からは外れて本末転倒かなと思う。
	現在の国家試験対策は基本的には過去問での対策がやっぱりメインになってる。過去問がないとなると新たに国家試験対策の方法を考え直さないといけない状況にはなるかなとは思う。
自己採点ができなくなる	自己採点をした結果を学校内で分析をかけていってどこの領域ができてないとか、その得点率とかを見ていって翌年に生かしていくんですけど、今回非公開になった場合に、合否だけではなく日本全国的にどういう問題が得点率が悪いよとかいうのが出てきたらありがたいなって。
	学生の自己採点とか振り返りが結果が出ないと分からないという形になると、それも手元に何も残らない状況で学生も不安になるのかなと思うので、試験を受ける時のことと、あと分からないというところで、どう問題に対応する力を教育していくのかというところが気になる点になります。
移行時のサポートの必要性	一番問題になるのは、試験の移行期の最初の学年だと思う。その人たちはすごい不公平だって、きっと感じる。2 年目は業者から過去問出る。切り替えに当たる人、あるいは前の学年で落ちそうで来年トライする人とか、要るじゃないですか。どの試験や制度でも変わる時に切り替えの人は移行措置を考えなきゃいけないのかもしれない。
問題の質保証の体制	ここで言ってるような質の保証っていう、第三者の目をどう入れるっていうあたりの仕掛けはやっぱりないと駄目だろうなと。あれだけ中でよくたたいて出したのに、不適切問題が出てくるって、そこの仕組みをどう整えるかは必要だと思う
試験対策のために必要な情報	もし仮に非公開になった場合にどんな準備が必要かとなると、今でも出題基準は出てるので、この範囲から出題されますということはある程度、明示されているが、その具体的な詳細の内容は公開されてない。提示されてないのでそれをまるっと提示する必要がある。本 1 冊ぐらい書いて、この中から絶対出ますというくらいしないと、過去問を公開しないことに国家試験出題側の耐えることができないかと思うので、それぐらいしないと駄目かな。
	そういうの（公式問題集）がいいと思います。恐らく国家試験の作問委員会、しかも厚労省とかが公式問題集みたいのを出して、それに準拠したものとかをもしかすると予備校さんとかもまねをして、そこから派生させてこういう予想問題みたいなのを出したりとかっていうのはあり得るかもしれない。 いずれにしてもどんな問題が出るかを公式に出す必要がある。現在の出題基準よりももっと丁寧に出さないと準備できないと思う。そこら辺は多分他の既に非公開を実施している多分、試験とかでやっていることをやる必要はあるっていうところ。
	アメリカとか非公開の CAT だけど、できているのは、検査センターが、「こういう問題って出せますよ」っていうのを公開、サンプル問題の公開はしているので、そういうふうにしておけば。
	試験問題の公開がないことで、どういうものが多く出題されてるかといった全体の傾向とか、難易度に関してのある程度の概要の情報公開はしてほしい
	練習問題だったり予想問題だったり、あと模擬試験という形でも何でもいいが、練習することができるようになっていると、過去問が公開されてない不安っていうのが学生も少なくって、準備もしやすいのではと思う。

表 34　シミュレーション教材等の活用状況（教員）

カテゴリ	具体的なコメント
LMS	学部では manaba で実習レポートの提出し、授業の資料を manaba に載せている。
	ウェブ教材とか提供しているのと、コロナ禍で病院に行けなかったときは看護学実習、私は基礎看護学なんですけど、看護学の実習を学内で模擬患者さんのシミュレーション情報を作って、それをウェブを介して学生が参加して、それで実習をするというような授業の組み立てをした。
	授業の案内は全て PC を通してとか、LMS を通して授業の案内。それから、成人看護Ⅰでは授業ごとに事前に LMS を通して授業資料を提供し、1 授業ごとに小テスト行って、確認テストと言われるものを行って。それも評価の中に入れるので、なしでは今はできない状況になっていると思う。
グループウェア	本校では Microsoft365、これを全員にアカウント配布してるので、全てのデータは学生と共有する、全てとは言わないけど、できる。なので、365 を使った授業とか、Teams で課題の提出や返却、授業資料の配布とか、Forms で授業アンケートの実施だとか、あとは授業前に「この動画見といてね」って学生に流して、それをもって授業に臨むみたいな事前学習的なことに使ったり
ビデオ会議	海外の人と直接会えない講義の時は Zoom
シミュレータ	人形を使って、あたかも人間のように技術を学ぶっていうことはしています。時々にコンテクストとか、その経過とかを設定して、それに合わせた看護の提供の仕方、看護というか技術になるが、主には。これに合わせた技術の提供の仕方を検討し、それをやることを試すこともあります。
	シミュレーションの教材、看護系の教材を利用して、多機能の人体モデルがしゃべったり、瞳孔反応が出たりして、心電図波形出たりっていう高機能シミュレータと、あと胸の音、胸部の異常音と肺、呼吸音の異常音とか聞けるシミュレータとかがあるので、そういうシミュレータを用いた演習とかを授業で計画することもある。それをウェブ経由で聞くこともある。
	コロナ中の実習に行けなかった時に、エルゼビアの vSim というのを買わせてもらって、ロールプレイングゲームみたいになってて、CAT ではないけれど。「今、患者さんの目の前にいました。何とかが起きてます。どうしますか」、ピコって選んで、最終的にすごくできると 100 点満点で、スコアが付くみたいな。あと、患者さんが「痛い」ってしゃべりだしたり。あまり時間がかかると「早くして」って怒ったりっというようなものは使ってみました。 （使ったのは）4 年生の総合実習。たったの 4 人とかのためにすごい高かった。だから、もう使えないってなったけれど。最初のころは、エルゼビアさんが、すごい使ってほしいから「どうぞ、無料で」みたいな感じで使わせてくれてて。でもあまりにも高かった。
	成人看護学ではこのコロナ禍になってから業者さんから SCENARIO をお借りしてた時期があったが、どうしても実習に行けなかったのでその SCENARIO を使って演習をしたり、本物の患者さんに見立てた事例として使ってた時期はあったが、学校のものではなくてリースしててもうお返ししたので今現在そういったものはない。
	シミュレーション教材を使ってて、僕、クリティカルケアの授業も担当しているんですけれども、いわゆる SCENARIO 君って商品名の人形があって呼吸音聞けたりとか、不整脈ってやると、モニターがあってモニターにその不整脈 Vf みたいなのが出たりっていう、血圧も数字として出たりっていうのを使っている。 なので使い方としては、まず学生に普段の事例を提示して、そこで「この患者さん、何観察するか考えときなさい」ってまず提示する。で、実際に SCENARIO 君が「人形が置いてある部屋に行って観察をしなさい」って。ちょっと事例、例えば不整脈になってたりとか、そういった仕掛けがあるが、それを用いて「じゃ、何観察した？」って。「何か普段と違ってたね。じゃ、それ何でかね。じゃ、どういうことアセスメントすればいいかな。何観察すればいいかな」っていうのを考えさせる授業をする。おおむね学生にはすごく好評で、やっぱり紙面の事例で展開するよりも、そういったものがあったほうがすごく分かりやすいっていうのは、学生の声として上がっている。
模擬(電子)カルテ	実際に救急カルテを使って、そこからどうやって情報を取ったらいいのかってことを学生自らちょっと学んでもらうように、そういった事例を作って、その事例をカルテに落とし込んでさらに。
動画教材	高齢者の状況であるとか、あとは認知機能が低下した人って、どういう状況なのかとかをやっぱり動画を使ったりとか、その動画を見てこういう状況になるんだっていうのを理解してもらったりというので、そういったメディアのものを使ったり。

学部定期試験以外での CBT 利用	大学院では使用している
	小テストに使用
コロナ流行時に定期テストで利用	テスト起動で。時間を決めて、よーいスタートで。見るなよって言っても見るかもしれなかったので結構問題量を多くして、見るといってもある程度調べないとできないような問題とかすごいいっぱい作った。ちゃんとばらついた。全員が多分見れる環境にいたと思うが、パソコンの前に座ってたので。で、見る、アクセス時間もちゃんと見て、2022 年ですかね、コロナ始まったときにやった。
	単純に manaba で選んでもらうっていうだけのやり取り。その時は、もう一斉にアクセスしてもらって、時間決めてというやり方。 その時、ちょっとやりづらいとか、ほんとに公平にできてるのかとか、カンニングとか採点とか、どうすんだとか。カンニングをしてもしょうがないって思って、何か見ていいことにはして、見れば答えられるっていう問題ではなくしてみたり。
	コロナ禍でも確かに試験はした。終講試験を PC で行ったが、これについてはやはり今すごく問題があるといったらおかしいが、まだ今課題が多いなと思う。試験についてはやはりコロナ禍においては学生の善意というか、良心を信じて試験は行ったが、各部屋、自らの責任においてやってもらったが、そこは本当に公平だったかどうかの確認は取れないという意味においては、だいぶ課題はあったと思う。 学生は家でやった。その時間に時間は合わせたけれども、LINE しながらでもできるから。そういう環境だったので少し課題は多いなと思う。集合型で国試でやられてるようなスタイルで考えるのであれば、それは問題ないと思う。

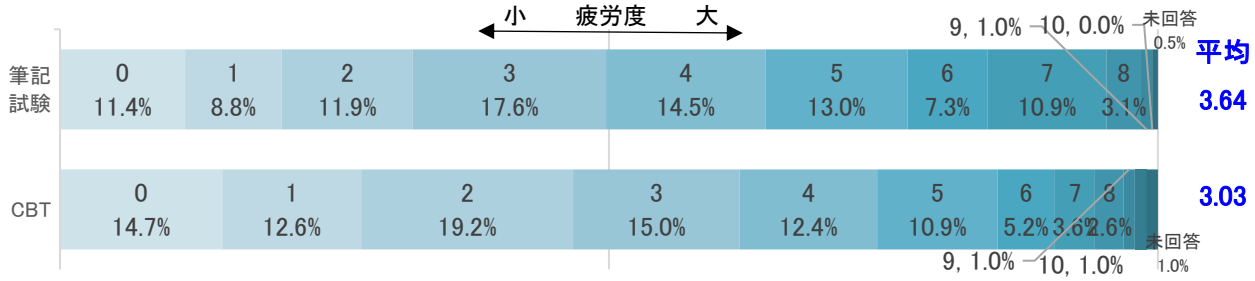
表 35　CBT に対応した教育の準備状況，移行に必要な期間についての意見（教員）

カテゴリ	具体的なコメント
現状の教育で対応できる	国際看護で出題するものが、すぐこの方式かと言われると、逆に、あまりないような気もする。例えば、今、テレビでやってる SDGs で、「SDGs は何個ゴールがありますか」とか。「前の MDGs とどう違いますか」とか。見せる問題、あんまりないなと思った。
	教育そのものが、すごくこれによって変わんなきゃいけないのかっていうのは、ちょっとイメージがあんまり、すいません、各論、今、担当してないので。
マルチメディア問題への対応	動画は授業でいっぱい使う必要があるだろう。対策を考えるのであれば、国家試験の一部が動画になるとすれば練習問題として授業の中にいっぱい入れてったほうがいい。(No.3)
	カリキュラム組めないからそれこそ動画、何パターンも模擬試験で準備してもらってそれもひたすらやるから。あとは動画で出る出題の範囲をある程度細かく出してもらうとした上で、教員が想定動画を作っておくか。そういう問題を売られるかもしれないが想定動画を作って準備しておくか。
学校・学生の経済的格差についての懸念	やっぱり明らかに動画を見る機会も少ないので、そういった動画を見る、見て判断する機会をつくるために、そういった動画ってすごい高いので、なかなかうちも手が出ない。なので、もう少し安価なもので、小規模の看護学校とかでも購入できるような、今でいう状況設定を動画化したものとかそういうものがあると、非常に実習の事前学習にも使えるかなと思うが、そういったものがあることで、もうちょっと読み取る力は上がっていくのかなと思うので、そういう動画類とか。
	今の学生の状況とか学校の経済状況とか、そういうところで懸念があるなとは考えていた。コンピュータを利用した試験の練習が例えば経済的に裕福な家庭とか、学校によって、事前に練習ができる、できないにばらつきができてしまうんじゃないかというのはある。特にうちの学校、市立の学校なので経済的に余裕がない家庭もあって、家にパソコンがなかったり、そういう子とかのことを考えるとちょっと。
	どういうふうに国家試験対策させればいいのとか、それこそ経済の格差が生まれるんじゃないかなと思う。経済格差っていうのは、結局予備校に通ってる人が強いみたいな。予備校はもう仕事なんだし、僕らも仕事なんですけど、予備校はもう全力で対策してきて、それなりの動画も準備してくると思う、きっとこうなった場合は。 けれども、そういった通うお金もないっていう子たちは、じゃ、何見て勉強すればいいのっていうのが路頭に迷っちゃうなっていうふうに思う。
CBT システムの操作のトレーニング	音声問題だったらここで、動画問題だったらここみたいな、操作のことは一定の慣れが必要なのかもと思って、音を聞く時、どこを押すんだっけって感じで迷ったりしたのがあったので、画面の操作のやり方は、慣れがちょっと必要かなと思う。
	パソコンを使った国試の問題ってなると、今回初めて問題を解かせてもらったが、選択肢の選び方だったり、次の問題に行き方だったり、そんなに難しくない画面ではあったが、国試の本番にこれを見たら、学生困るかなと思った。なので、国試の問い方に合わせた問題の答え方の練習とか、パソコンの操作の仕方を事前に練習しておいて緊張度を下げたほうがいいと思った。
不足している内容準備期間	それは、もう圧倒的に演習の時間が足らないなと思っている。やっぱり知識ベースというか、出題範囲。国試の授業を組み立てる上でも、国試の範囲っていうのは一応意識はしますので、それらもカバーしようと考えると、基本的な知識や看護を行うために必要な知識の部分というものの確認する時間が、授業時間の中でも結構占める割合が高いので。やはりこの判断力を問うような問題を解けるようにしていくためには、さらにもっと演習時間を増やすっていう必然性はあるんじゃないかと思っています。
	ああいう質問、もちろん臨床に行くと恐らく求められる力なんだなって思うので、もう少し実習の中で、すごい大変だが、今は看護課程中心にやっている、ほんとにその場その場での判断とか状況を読むっていうところで、今いわれてる臨床判断とか推論とかをもうちょっと実習の中で強化していかないと、あれには対応できないなって思ったのと、そこに、実習にそういうものを入れるんであれば、学内での講義とかももう少しいろんな知識をつなぐような、もちろん今もやってはいるが、なかなか成果が出てなくてあれなんですが、例えば解剖とかの知識、病態とかの知識をつないで、看護とつなぐところを、授業の中でもう少し意図的にやっていかないと、実習での臨床判断とか推論とかの実習には持っていけないんだろうなと思うので、全体的にもう一回カリキュラムを見直さなきゃいけないのかなと思った。

その場面を見て状況をつかむ能力とか、あとこれがどうなるんだろうって予測するような能力っていうのは、やっぱり教育内容としては必要なのかなと思うが、現状それはあまりできていない.
国試がコンピュータを使ったものに変わる場合、学生の国家試験対策の教育面での準備は、3年あればいい。
学びのほうに関しては、やっぱり 2 年か 3 年あるほうが良いのかなって、学部が入って、せいぜい 2 年生ぐらいにはそういうことになるんだっていうのが分かりつつ、準備をするぐらいの期間があるほうがいいのかなと思う。 ただ、もちろん来年なると言っても、やればするんじゃないかとは思うが、でもちょっと乱暴な感じ。あと、機器を使うほうの準備は、やっぱり 1 年くらいあればいいのかなと。専門学校とかはそういう練習するかもしれない。
入学のときから必要かと思うので 4 年間ぐらいは必要と思う。大学 1 年に入学したときから今うちの学校でも国家試験を意識させて 4 年生の 2 月に受ける国家試験の合格率が上がるようにとやっているので、それを考えるとやっぱり 4 年前かなと思う。2 年生になってから 3 年後、CBT だよって言われても、ええーみたいな感じなので。やっぱり入学時からあなたたちは CBT ですっていうことを分かってたほうがいいと思うと、その子たちが入学する前から教員は準備しなきゃいけないとなると、学生としての準備期間は 4 年間一緒だと思います。 教員としては 5 年前ぐらいじゃないと。新入学生の国試対策を準備するのには4年かかると思うと 5 年ぐらい前から告知されないと準備できない。
3 年かな。と思ったのは入学した学生が、「あなたが卒業する時は、CBT なんです」っていうのを知って、テストしてもらいたいから。少なくとも 3 で。5 年ぐらいはかかるのかなって思う。ほんとに、何の根拠もなく。 それこそ、パソコンに慣れるみたいなところから始まることになったら、「いい、普段から、パソコンで、もの書くのよ」とかから始めなきゃいけない人たちっている。
結局、彼ら（国試向けに）1 年しか勉強しないので。問う知識は（紙と CBT で）同じですよね。だからといって、例えば、今までだったらシナリオだけでずっとやってたものが、画像になる。だから、演習も全部できるだけ画像でやってとか、だから 4 年必要とか、そういうことはあんまり思わない。少なくとも試験に出るようなことは、多分全て教材、電子教材になってるので、それを使って勉強するっていう感じになると思う。1 年前だとあれかもしれないんで 2 年前ですかね。2 年前にアナウンスして、学生たちは 1 年前から心して試験対策するという、そんな感じかなと思います。
3 年間で国家試験を受けるようになるので、入学時からこういう試験になってるということで 3 年間準備をすればというとこかなと思う。
授業計画とか計画を立てるための 1 年前になると、その受ける学年の在籍数プラス 1 年は必要なんじゃないのかなと思います。そうしないと、学生にきちんとした教育プログラムで学習するというのは難しくなるんじゃないのかなと思います。

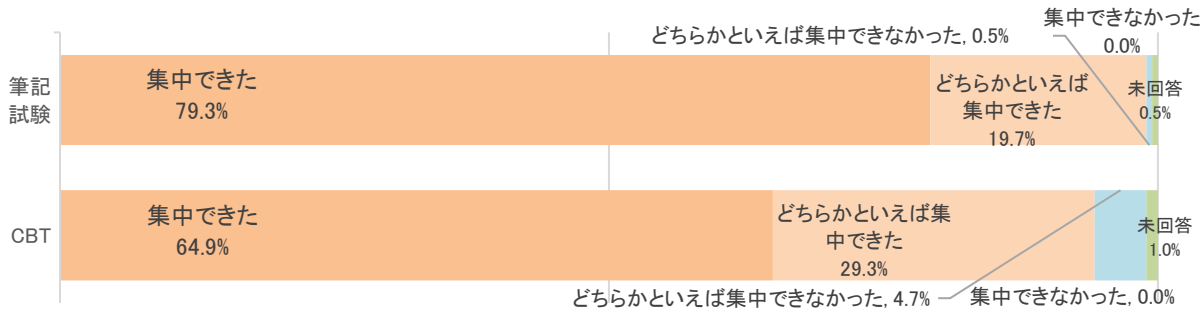
図 17　CBT による試験実施の課題に関する調査　学生対象調査　解答後アンケート集計結果

Q1 現在、あなたはどの程度疲れを感じていますか

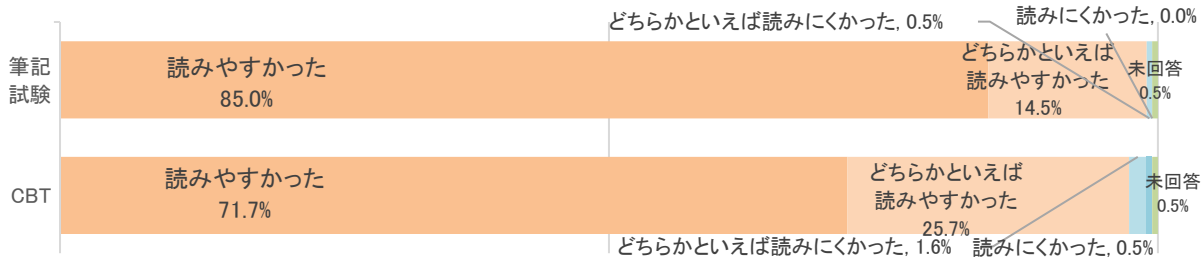


p＝0.005　有意差あり（Mann-Whitney 検定）

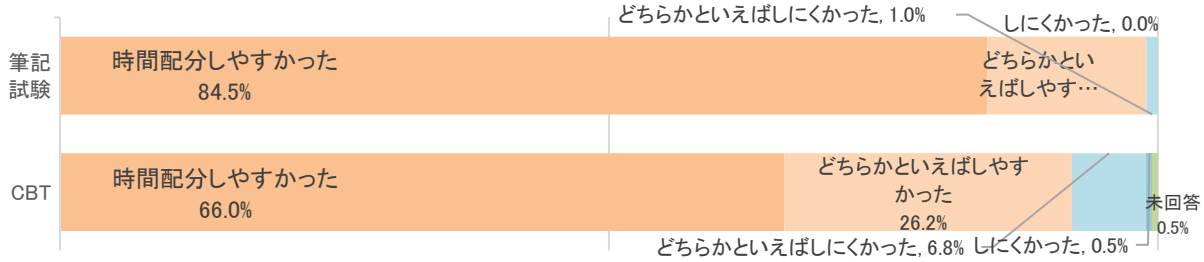
Q2-1 解答に集中できましたか



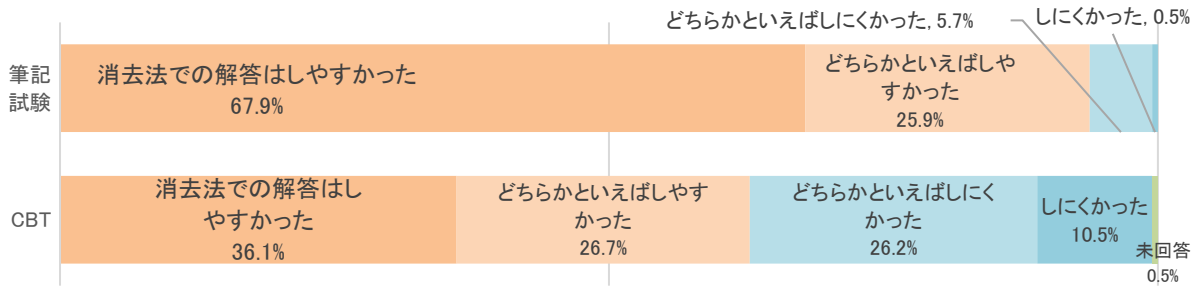
Q2-2 文字は読みやすかったですか



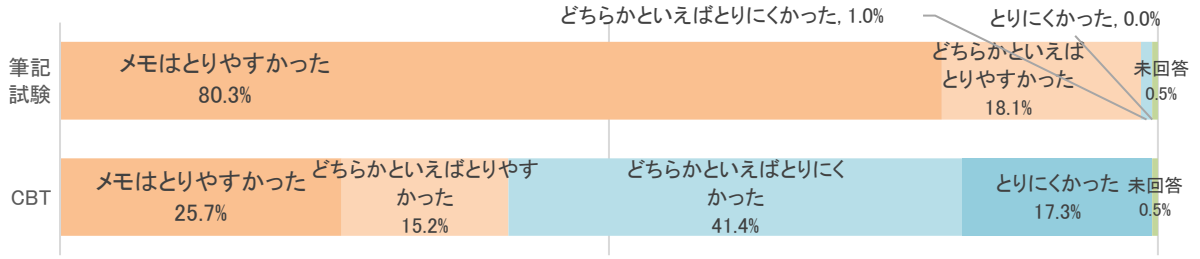
Q2-3 時間配分はしやすかったですか



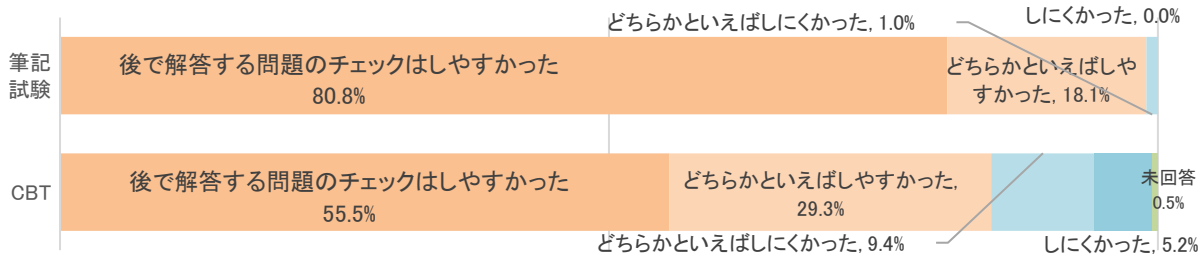
Q2-4 消去法での解答はしやすかったですか



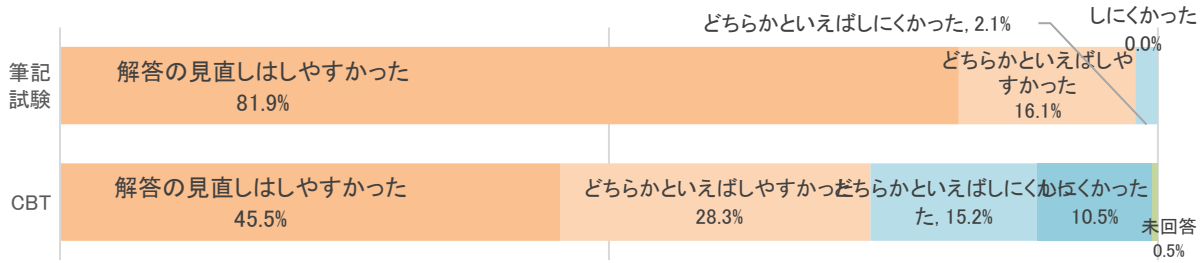
Q2-5 メモはとりやすかったですか



Q2-6 後で解答する問題のチェックはしやすかったですか



Q2-7 解答の見直しはしやすかったですか



Ⅱ．研究成果の刊行に関する一覧表

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
Yusei Kido, Hiroaki Yamada, Takenobu Tokunaga, Rika Kimura, Yuriko Miura, Yumi Sakyo 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	Volume 1	821-829	2024年
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		1074-1079	2025年
Yusei Kido, Hiroaki Yamada, Takenobu Tokunaga, Rika Kimura, Yuriko Miura, Yumi Sakyo 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	Volume 1	754-764	2025年

以上