文部科学省委託事業

学力調査を活用した専門的な課題分析に関する調査研究

（全国学力・学習状況調査の CBT 化に向けた試行・検証）

A.　CBT 記述式答案の採点に関する試行・検証

最終報告書

令和７年３月

株式会社内田洋行

教育総合研究所

目次

1 調査研究の目的と背景

令和７年度以降段階的に CBT への移行が予定される、全国学力・学習状況調査（以下、「全国学調」と称する。）の実施に関し、悉皆調査において CBT を活用する意義の１つとして、「より効率的な採点の実現」が示されており、「記述式問題に対する児童生徒の答案も機械可読なデータとなることから、機械採点の導入による採点の効率化を検討する（本事業仕様書より抜粋）」ことが求められています。

令和７年度の全国学調仕様書においても、「中学校理科以外については調査基準日から概ね６週間以内までに、中学校理科については項目反応理論（Item Response Theory,IRT）による分析とそれに基づいた結果返却を可能とするため、概ね４週間以内に、採点を完了すること」「適切な環境下で機械採点システムを導入し活用することなどを認める」と、CBT 対象教科に対する採点期間の短縮と、それに伴う機械採点の活用が想定されています。

このため、本事業では、機械採点の導入による採点の効率化の程度を検証し、機械採点の導入が、仕様に求められた採点期間短縮の一助となることを示すため、機械採点（以降「自動採点プログラム」とも称する。）を用いた場合の処理速度や精度について測定・推計することで、採点完了までの期間短縮に関する試行・検証を行います。

2 現行の全国学力・学習状況調査の採点業務工程

本項では、令和６年度まで人手による採点業務の中で行われていた工程について確認します。

令和７年度以降 CBT や機械採点プログラムが導入される場合でも、人間が関与する工程については概ねこれに準じた工程がなされると推測されます。

2.1 現行の採点業務の仕様

令和６年度全国学調仕様書では、採点業務及び採点者、採点監督者に関して、それぞれ下の通り示されていました。

過年度の人手による採点業務工程はこれら仕様に基づきシステム等が構築されており、令和７年度の中学校理科の調査にあたって自動採点プログラムを併用する場合にも、人手による採点業務では、令和６年度仕様に準じた運用がなされると想定されます。

項目	仕様書の記載
採点方法 採点者 (仕様書より抜粋)	本体調査及び経年調査における短答式・記述式・口述式の問題については、２回採点（１つの問題を採点基準に基づき独立した２人以上の採点者が採点。）による採点を行うこと（無解答の解答類型に相当する解答については、機械による採点２回に代えることも可能。）。短答式・記述式・口述式の問題については、それぞれ適切な能力を有する採点者（又は機械）による採点を行うこと。 機械による採点については、採点品質を担保するために、採点監督者等による品質チェックを適切に行うこと。
採点監督者 (仕様書より抜粋)	採点のミスやぶれを無くすため、採点監督者が採点結果の確認や不一致答案（複数の採点者による採点結果が一致しなかった解答）の採点を行うなどすること。これらにより、本体調査における短答式、記述式及び経年調査における短答式、記述式、口述式の問題について、採点基準に基づき独立して採点した２人の採点者（又は機械）の採点結果が一致するまで採点を行うこと。
プレ採点 (仕様書より抜粋)	採点作業の初期段階においては、採点結果の点検を充実させ、事前に想定し得なかった解答例を中心に、国立教育政策研究所と作業方針に係る調整を行うなど一連の採点作業を円滑に実施するための必要な調整を行うこと。また、採点作業中に随時、国立教育政策研究所と作業方針に係る調整を行うこと。特に、調査実施後採点作業が始まる前の段階で、採点基準に基づき正確な採点が行われるように、実解答を基に採点マニュアルについて国立教育政策研究所と十分に協議を行うこと。
品質チェック (仕様書より抜粋)	採点作業中に、随時、採点結果の品質チェック及び改善を行い、採点の質を確保すること。 本体調査における短答式の問題は、問題ごとに 5,000 件を抽出した際に採点結果の正誤の誤りが１件も無くなるまで採点を行うこと。 本体調査における記述式の問題は、問題ごとに 1,000 件を抽出した際に採点結果の正誤の誤りが１件も無くなるまで採点を行うこと。 本体調査及び経年調査ともに、正誤の誤りが無いことが確認されるまで、繰り返し採点結果の品質チェックを行った上で国立教育政策研究所において、品質チェックが適切に行われたことを確認することによって、採点作業の完了とする。

2.2 現行の人手による採点業務工程

現行の人手による採点業務の流れについては、採点マニュアルを随時更新しながら、下の表に示すような流れで行われると想定されます。

工程	作業の概要	主となる作業者
独立２回採点	マニュアルに基づき、異なる採点者２名が、それぞれ独立して判断を行い、採点結果をシステムに入力する。	一般採点者
保留・不一致解答の判断	一般採点者が判断に迷った解答や、２名の判断が一致しなかった解答に対し、上位採点者が採点結果を付与する。	上位採点者
品質確認	採点結果が付与された解答を再度目視し、判断が誤っている解答、疑わしい解答について検出する。	一般採点者・上位採点者
検出された解答の再確認	上の品質確認で検出された解答に対し、再度上位採点者により内容を確認して判断を行う。	上位採点者
未分類解答の最終判断	上位採点者でも判断に迷った解答に対しては、教科責任者が個々に最終的な判断を行う。	教科責任者

機械採点では、ここで示した工程のうち「独立２回採点」で一般採点者が行う作業の１回以上を置き換えると想定され、置き換えた場合の効率化について本報告書の中で推計します。

2.3 現行の採点スケジュールの概略

令和６年度全国学調仕様書では、採点期間と結果集計期間について、それぞれ下の通り示されていました。

項目	仕様書の記載(抜粋)
採点期間 (仕様書より抜粋)	国立教育政策研究所の作成する解答類型及びその分類の考え方に基づき、迅速、正確かつ確実に採点を行うための仕組みを構築すること。本体調査は調査日から概ね６週間以内までに採点を完了すること。
結果集計期間 (仕様書より抜粋)	採点結果及び質問紙調査の回答を迅速かつ正確に集計し、集計結果から、本体調査については一覧表に示す資料を作成し、文部科学省の点検を経た上で、採点作業終了後、おおむね３週間以内に納品すること。

これをもとに、過年度の人手による採点業務工程では、概ね次に示す日程で進行していると想定されます。

時期	実施工程の概略
～調査実施日	・採点基準に基づく採点を行う仕組みの構築 ・採点マニュアルの作成 ・採点システム、会場等の構築 ・採点者の採用
調査実施日 ～採点作業開始まで	・採点者の研修 ・上位採点者による先行採点
採点作業開始 ～調査実施日から６週間	・一般採点者による採点作業 ・一般採点者、上位採点者による品質向上作業 ・国立教育政策研究所による検収確認 ・採点完了、結果集計の開始
～採点終了後３週間 （調査実施日から概ね９週間）	・集計完了、納品

自動採点では、一般採点者が行っている２回の採点のうち１回以上を置き換えることが期待されます。したがって上の表で一般採点者が採点作業を行う「採点作業開始～調査実施日から６週間」にあたる日程をどの程度効率化できるかを、本報告書のなかで推計します。

3 自動採点の活用の検討

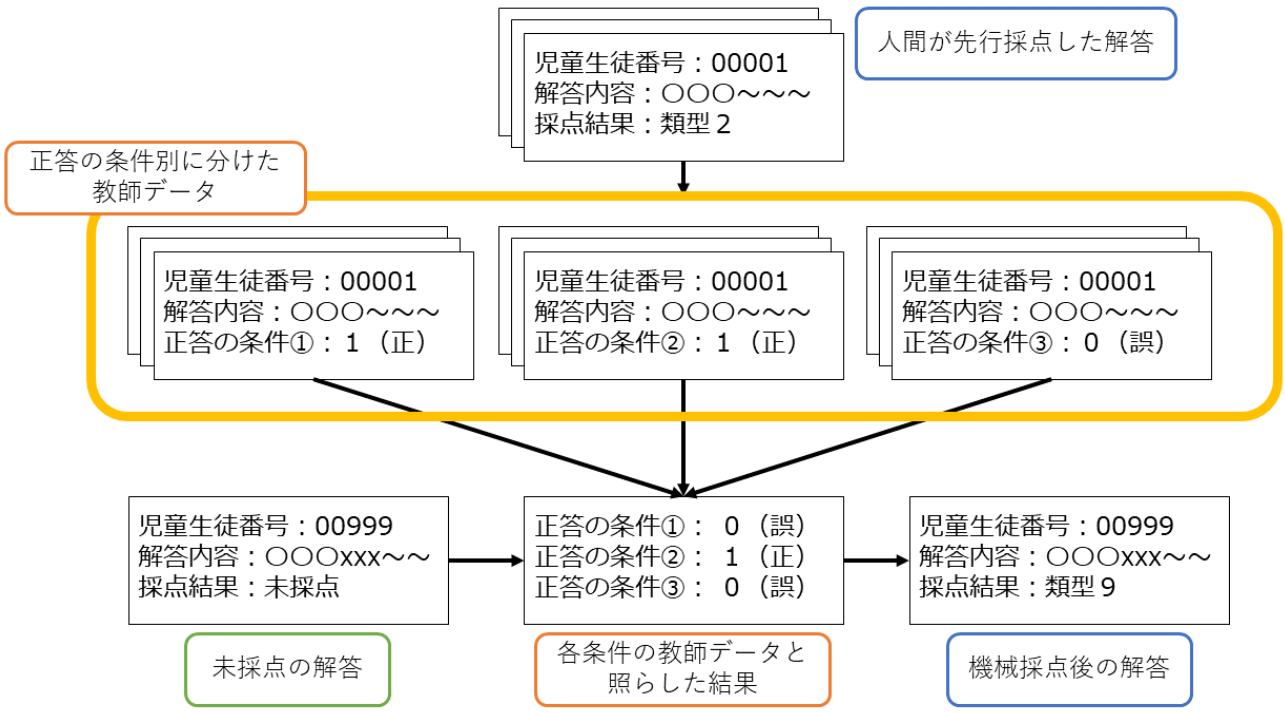
3.1 使用した自動採点プログラムの概略

3.1.1 本事業で用いた自動採点手法の基本動作

全国学調で出題される記述式問題は、原則として１つ以上設定された「正答の条件」のそれぞれを満たすかどうかで、類型および正誤が分類されます。つまり、自動採点プログラムが正答の条件別に正誤を判断することができれば、その結果を組み合わせることによって正しい類型を出力することが可能です。

このため、今回の試行検証で使用する２つのプログラムでは下の図のように、人間が先行採点することで収集した解答と類型の組み合わせを、まず正答の条件別の正誤（正：１または誤：０）として変換し、正答の条件ごとに教師データとしてまとめます。

その後、新たに入力する未採点の解答に対して、それぞれの正答の条件を満たすか否かの二値分類を行い、二値分類の結果を結合することで、最終的な類型を出力するプログラムを作成、使用しています。



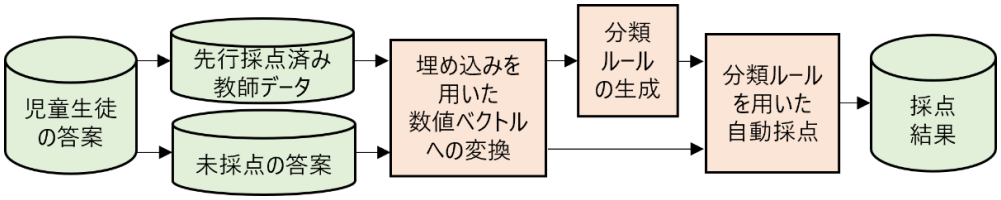
3.1.2 本事業で用いたプログラムの概略

本項では、本事業で用いたプログラムの処理とハードウェアの構成について、その概略を示します。

なお、クラウド環境に対するセキュリティについては、国内サーバーのみに構築を限定し、指定した IP アドレスのみの接続として、本事業関係者以外のアクセスを制限したほか、解答の文章が他の Web サービス等に学習されないよう配慮した処理を行いました。

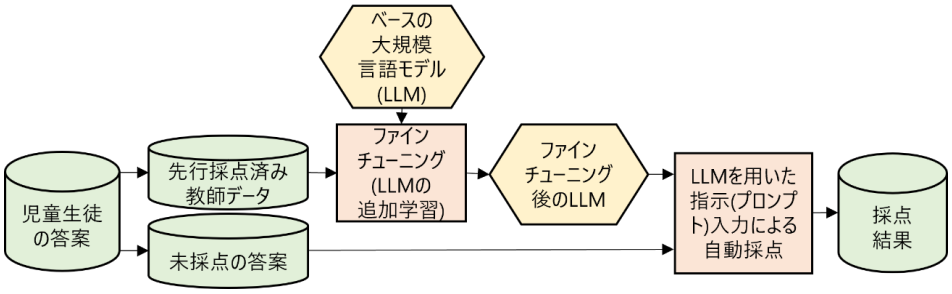
機械学習手法を用いた分類（以降「プログラム A」と称する。）の概略

児童生徒の答案を、埋め込みモデルを用いて数値ベクトルに変換し、先行採点の結果を教師データとして、答案が持つ数値ベクトルをもとに、機械学習により正答の条件の適否や解答類型（以下、「類型」と称する。）に分類するためのルールを生成し、その分類モデルに基づいて未採点の答案の分類・自動採点を行います。



生成 AI の原理を用いた分類（以降「プログラム B」と称する。）の概略

生成 AI の基盤として用いられる大規模言語モデル（LLM）に対し、答案と出力すべき類型の組み合わせを追加で学習（ファインチューニング）させて、特定の設問を分類することに特化した大規模言語モデルを生成します。ファインチューニング後の大規模言語モデルに対し、類型を分類するための指示を入力し、指示に従って未採点の答案の分類・自動採点を行います。



3.1.3 ２種類以上の自動採点プログラムを稼働させる場合の留意点

前項に示したような自動採点プログラムが複数種類存在する場合、複数種類の自動採点プログラムを採点業務の中で同時に稼働させることも考えられます。

本事業のなかで複数種類のプログラムを稼働させた際に生じた課題をもとに、悉皆調査内で複数種類の自動採点プログラム稼働させた場合に留意点を下の表に示します。

観点	留意点
総計算量の増大	各解答に対して複数種類の自動採点プログラムそれぞれで計算を行うことで、機械学習等を行うための計算量が増加する。 計算量が増加するほど GPU 等の計算資源の確保が求められ、併せて必要な費用が増加すると考えられるため、複数種類の自動採点プログラムを稼働させた際の効率化の程度が、費用の増加に見合うものかどうかを考慮する必要がある。
出力結果の統合	複数種類の自動採点プログラムを稼働させる場合、使用するアルゴリズムや言語モデルの違いによって、それぞれの自動採点プログラムが出力する採点結果が異なる可能性がある。 そのため、それぞれの自動採点プログラムが出力した採点結果をどのように統合するかや、それぞれの自動採点プログラムの結果が食い違った際にどのように結果を採用するかどうかを考慮する必要がある。
自動採点を行えない場合の対応策	採点作業の効率化を求めて自動採点プログラムを稼働させる場合、クラウドサービスの障害など何らかの事情でプログラムが稼働できなくなった際に、採点者による採点作業工数の増大リスクについて考慮する必要がある。 これはプログラムを１種類のみ稼働させる場合にも留意すべき観点ではあるが、複数種類のプログラムを稼働させる場合にはより少ない採点者数での作業実施が予想され、なおかつ複数稼働するプログラムのうち一部が停止する可能性が生じる分、よりリスクが増大すると考えられる。

3.2 実施した採点の方法

3.2.1 検証対象問題の選定

詳細な分析を行う問題について、仕様に基づき、令和６年度全国学調経年変化分析調査問題の中から、学力調査室と国立教育政策研究所との協議により選定しました。

選定にあたっては、経年変化分析調査の実施教科に理科が含まれないことから、主に次の表に示す４つの観点により選定しました。

観点	内容
中学校理科への推定	特に令和７年度に CBT で実施予定の、中学校理科で出題されるような問題形式に対して、本事業の測定結果をもとにした推定ができるような題材や出題形式であること。
解答内容の多様性	解答内容に多様性があり、同じ類型でも典型的な解答から判断に迷うものまで、幅広い内容の収集が期待できること。
採点の難易度	文章量や正答の条件の数などの点から、正しい類型に分類するための難易度が高く、選定した問題で自動採点プログラムが機能すれば、他の問題でも機能することが期待できること。
典型的な出題形式	過去にその教科で複数回出題されている出題形式や題材で、今後も同様の形式による出題が期待できること。

これらの観点に基づき、全国学調受託事業者と連携し、小学校国語と小学校算数から各１問ずつを、詳細な分析を行う問題として選定しました。これ以外の経年変化分析調査問題については、本年度末までに測定を行うこととしました。

3.2.2 自動採点の流れ・人手による採点との差異

人手による採点作業では、項番 2.2 で示したような工程で採点が進行していきますが、本試行検証で用いる自動採点プログラムでは、下の表に示すとおり採点結果の出力を得ています。

使用プログラムの種類	工程の概略
1 種類による採点 （独立 2 回採点の一方の代わりとすることを想定）	・プログラム A または B により、各正答の条件の正誤を判断する。 ・全正答の条件の正誤判断後、その組み合わせに対応した類型を出力する。 ・文章の内容を問わず、全ての解答に対して判断結果を出力する。

このため、過年度を含め人手により行われている採点の工程とは、下の表に示すような差異があることに留意が必要です。

人手による採点との差異	留意点
保留の有無	人手による採点の場合、採点マニュアルの記載に照らしても判断に迷うものや、解答主旨が不明なものに対しては「保留」と分類して、上位採点者等に判断を仰ぐのが通常です。 一方で、本試行検証で使用したプログラムでは、各正答の条件に対する判断結果として、どのような場合でも正誤（1 or 0）いずれかを出力するように構築しています。
採点時に確認する正答の条件	人手による採点の場合、落書きなど判断に値しない解答や、条件を複数満たさず判断途中で誤答が確定した解答などに対して、各正答の条件に対する判断の一部または全部を省略して採点する場合があります。 一方で、本試行検証で使用したプログラムでは、常にすべての解答に対しすべての正答の条件を判断して類型を得るため、測定にあたっては、人手による採点結果を一度正答の条件単位に変換した上で教師データと扱っています。

3.3 実施した自動採点の精度測定

3.3.1 精度測定にあたっての条件設定

自動採点プログラムに必要な教師データの抽出や、精度の測定手法については、次の表に示す通り行いました。

観点	内容	根拠・背景
取り扱うデータ	上位採点者１名のみで採点された段階の採点結果データを全国学調受託事業者より受領して、分類モデルの作成および自動採点の精度測定を行った。	経年変化分析調査の業務実施工程に影響を与えないようにするため
測定の流れ	特定の乱数シード（乱数生成の初期値）により、１回採点済みのデータからランダムサンプリングを行い、教師データを抽出して分類モデルを作成する。 教師データとして抽出した以外の解答は、検証データとして未採点の状態から改めて自動採点を行い、人手による採点結果との比較を行う。	乱数シードを固定することで、プログラムＡ・Ｂの間で同一の教師データと検証データを用いて測定、比較を行う
測定回数	５つの異なる乱数シードを用い、上記のランダムサンプリング・分類モデル作成・自動採点をそれぞれ５回測定する。	本事業の期間内で処理の完了が見込まれ、なおかつ採点精度の範囲を一定程度示すことができると考えられる回数
教師データとして用いる解答件数	１回の測定あたり 500 件を教師データとして抽出する。	どのような出題内容でも概ね１日以内で上位採点者により先行採点を行い、正しい採点結果の収集が見込まれる件数

3.3.2 自動採点プログラムの採点精度測定結果

本項では、各自動採点プログラムによる測定結果を示します。

プログラム A（機械学習手法を用いた分類）の採点精度

①：小学校国語の採点精度

小学校国語の選定問題について、自動採点を行った結果を下に示します。

①-1-1：正誤、類型、各条件別の、人手による採点結果との一致率

精度	正誤一致	類型一致	条件①	条件②	条件③
平均値	87.54%	75.35%	89.03%	84.86%	89.52%
最大値	88.16%	76.19%	89.50%	85.82%	90.65%
最小値	86.64%	74.34%	88.49%	83.67%	89.04%

※精度は小数第３位を四捨五入。以降同様。

①-1-2：人手による採点結果と、自動採点結果の組み合わせ件数（5 回の合計）

列：人間の採点結果 行：自動採点結果	◎類型 1	×類型 2	×類型 3	×類型 4	×類型 9	自動採点の類型一致率
◎類型 1	3669	63	605	642	154	71.48%
×類型 2	1	0	15	4	5	0.00%
×類型 3	99	7	1423	25	243	79.19%
×類型 4	714	11	136	7613	1548	75.96%
×類型 9	7	5	65	170	1111	81.81%

※ランダム抽出５回の合計のため、各行列の合計が５の倍数になるとは限らない。以降同様。

特に自動採点の結果類型２に分類された解答の中に、人間が類型２と判断した解答が１件も含まれない結果となりました。これは類型２と関連する正答の条件③が、解答の内容によらず入力された文字数のみによって判断されることから、文章の意味内容を解釈して分類する自動採点プログラムでは、内容が正しく字数の過不足があるような解答について、内容が正しいことだけを認識して類型を分類してしまった可能性が示唆されました。

正答の条件③については、関数などを利用したルールベースによる分類のほうが望ましい結果となることが推測されたため、文字数をルールベースでカウントした上で、内容の判断結果によって

分類するよう調整して再度分類しました。その場合の精度を次に示します。

①-2-1：字数をルールベースで自動化した場合の、人手による採点結果との一致率

精度	正誤一致	類型一致	条件①	条件②
平均値	88.44%	75.70%	89.03%	84.86%
最大値	89.06%	76.57%	89.50%	85.82%
最小値	87.76%	74.69%	88.49%	83.67%

①-2-2：字数をルールベースで自動化した場合の、人間と自動採点の結果組み合わせ（5 回の合計）

列：人間の採点結果 行：自動採点結果	◎類型 1	×類型 2	×類型 3	×類型 4	×類型 9	自動採点の類型一致率
◎類型 1	3670	0	560	615	125	73.84%
×類型 2	0	63	60	31	34	33.51%
×類型 3	99	7	1423	25	243	79.19%
×類型 4	714	11	136	7613	1548	75.96%
×類型 9	7	5	65	170	1111	81.81%

内容部分の分類結果に変化はありませんが、字数により誤答となる解答が正しく分類されることによって、全体としての分類精度は向上しています。

字数の判断に代表される、適否の判断に厳密なルールが設定できる正答の条件に対しては、ルールベースによる分類を併用して形式的な誤りを検出することにより、誤答となるはずの解答を正答と誤って分類する可能性を低くすることが期待できます。

そのうえで、類型 2 に対する採点精度が低かった理由として、記述が少なく字数不足となる解答は、どのような内容かを解釈する材料に乏しい場合が多く、誤答の中でどの類型となるか、内容判断が困難であったためと推定されます。

②：小学校算数の採点精度

小学校算数の対象問題について、自動採点を行った結果を下に示します。

②-1-1：正誤、類型、各条件別の、人手による採点結果との一致率

精度	正誤一致	類型一致	条件①	条件②	条件③	条件④	条件⑤
平均値	95.75%	78.59%	97.81%	95.94%	85.53%	97.79%	97.48%
最大値	96.14%	79.50%	98.13%	96.02%	86.40%	98.41%	97.76%
最小値	95.44%	77.79%	97.61%	95.83%	84.74%	97.46%	97.12%

②-1-1：人間による採点結果と、自動採点結果の組み合わせ件数（５回の合計）

列：人間 行：自動	◎ 類型１	○ 類型２	○ 類型３	× 類型４	× 類型５	× 類型６	× 類型７	× 類型８	× 類型９	自動採点の 類型一致率
◎類型１	52	0	2	0	6	0	0	2	1	82.54%
○類型２	0	0	0	0	0	0	0	0	0	該当なし
○類型３	209	0	80	6	43	4	0	4	5	22.79%
×類型４	9	4	209	228	52	1	1	0	1	45.15%
×類型５	108	1	277	381	994	0	5	11	4	55.81%
×類型６	0	0	0	5	0	3708	836	0	17	81.21%
×類型７	0	0	0	0	6	994	6003	0	34	85.31%
×類型８	0	0	0	0	0	0	0	1629	6	99.63%
×類型９	6	0	8	11	66	6	20	133	132	34.82%

正誤分類の結果や、各正答の条件の採点精度は非常に高くなっている一方で、最終的な類型一致率については、複数の類型で精度が非常に低い結果となりました。

この背景として、正答の条件③を除いた各条件の反応率が低く、本設問の正答率が非常に低かったことが挙げられます。反応率が極端に低い条件に対してランダムサンプリングで抽出した教師データを学習させると、全ての解答を誤答と判断することで精度の高い結果が得られるように見える分類モデルが作成されやすく、正答の条件を一部満たすような解答を正しく分類できなかったと推測されます。

このため、先行採点の結果などから反応率や正答率が極端に低いことが予想される場合、教師データの内容に偏りが生じないよう、教師データの選択に配慮する必要があったと考えられます。これは正答率が極端に高い場合でも同様のことがいえます。

加えて、本設問では、誤った選択肢を選んだ場合、記述の内容によらず類型が定まる場合が存在しました。類型６以降がこれに該当し、選択肢の分類が正しければ内容部分の解釈を一部誤っていても同じ類型に分類されるため、これも見かけ上の分類精度を高めることにつながりました。

また、誤った選択肢で類型が限定される解答の場合は、小学校国語の場合と同様に、ルールベースによる自動分類の方が高い精度となることが考えられます。

これらの要因を踏まえ、3,764 件のうち正しい選択肢を選んだ 637 件のみを測定対象（教師データ 500 件、検証データ 137 件）として、再度自動採点を行いました。

②-2-1：正しい選択肢を選んだ解答のみを対象にした、人手による採点結果との一致率

精度	正誤一致	類型一致	条件②	条件③	条件④	条件⑤
平均値	85.99%	71.82%	84.09%	92.26%	93.87%	91.97%
最大値	89.78%	76.64%	85.40%	94.16%	95.62%	94.16%
最小値	82.48%	69.34%	83.21%	89.78%	92.70%	89.05%

②-2-2：正しい選択肢を選んだ解答のみを対象とした場合の結果組み合わせ（５回の合計）

列：人間の採点結果 行：自動採点結果	◎類型１	○類型２	○類型３	×類型４	×類型５	自動採点の 類型一致率
◎類型１	78	0	7	0	10	82.11%
○類型２	0	0	0	0	0	該当なし
○類型３	12	1	91	8	9	75.21%
×類型４	5	0	39	101	17	62.35%
×類型５	7	0	18	60	222	72.31%

解答全体を対象とした測定よりも精度が低下したように見えますが、自動採点プログラムが持つ解釈能力の測定結果としては、こちらの結果の方がより実際に即した値と考えられます。

小学校国語と比較すると、各条件単体の精度としては国語と大きく差のない精度で分類できている一方で、類型分類に必要な条件の数が２つから４つに増えていることで、全ての条件を組み合わせた最終的な類型の分類精度が、小学校国語よりも低くなったと推測されます。

プログラム B（生成 AI の原理を用いた分類）の採点精度

次に、プログラム B を用いた自動採点の結果を示します。

なお、プログラム A の測定時に確認された内容を踏まえ、ルールベースで判断可能な条件については割愛し、記述内容に着目した判断精度を測定するものとしました。

①小学校国語の採点精度

①-2-1：字数をルールベースで自動化した場合の、人手による採点結果との一致率

精度	正誤一致	類型一致	条件①	条件②
平均値	83.94%	67.88%	83.64%	84.15%
最大値	89.42%	73.58%	89.20%	90.07%
最小値	78.35%	59.23%	78.07%	78.89%

①-2-2：字数をルールベースで自動化した場合の、人間と自動採点の結果組み合わせ（5 回の合計）

列：人間の採点結果 行：自動採点結果	◎類型 1	×類型 2	×類型 3	×類型 4	×類型 9	自動採点の類型一致率
◎類型 1	2615	0	547	403	119	70.98%
×類型 2	0	30	0	0	0	100.00%
×類型 3	180	7	1200	23	163	76.29%
×類型 4	1606	43	256	7300	1478	68.33%
×類型 9	89	6	241	728	1301	55.01%

精度の最大値はプログラム A よりもわずかに小さい程度でしたが、精度の最小がプログラム A よりも大幅に落ち込んだことにより、平均としての精度が下がる結果となりました。

最小値が落ち込んだ原因として、ほとんど全ての文章を誤答として判断するような出力により、正答となるはずの文章が大量に誤答に分類されており、正確な原因は不明ながら、正答と結びついた教師データが少なかったことで、言語モデルに対するファインチューニングが誤答の側に極端に働いてしまったことが考えられます。

プログラム A のように数値化したものに対する分類手法と異なり、プログラム B のように生成 AI の原理による自動採点では、文章として出力させる内容を完全にコントロールすることができないため、事前の教師データを十分確保するなどの出力に異常を起こさない工夫や、出力に異常が見られた場合の対応などについても検討が必要と考えられます。

②小学校算数の採点精度

②-2-1：正しい選択肢を選んだ解答のみを対象にした、人手による採点結果との一致率

精度	正誤一致	類型一致	条件②	条件③	条件④	条件⑤
平均値	79.56%	53.43%	70.95%	70.36%	90.66%	90.95%
最大値	86.13%	62.04%	74.45%	75.91%	93.43%	95.62%
最小値	75.18%	48.18%	67.15%	64.96%	83.21%	88.32%

②-2-2：正しい選択肢を選んだ解答のみを対象とした場合の結果組み合わせ（５回の合計）

列：人間の採点結果 行：自動採点結果	◎類型１	○類型２	○類型３	×類型４	×類型５	自動採点の 類型一致率
◎類型１	80	1	21	1	15	67.80%
○類型２	0	0	0	0	0	該当なし
○類型３	14	0	32	8	6	53.33%
×類型４	0	0	28	32	15	42.67%
×類型５	8	0	74	128	222	51.39%

プログラム A と同様、算数では正答の条件が４つあることから、類型一致の精度の落ち込み方が国語よりも顕著に現れています。

また、すべての文章を誤答として判断するような出力も複数回確認され、特に小学校算数では、国語と比べてさらに各条件の反応率が低いことから、より極端なファインチューイング結果となってしまったことが考えられます。

3.3.3 その他の条件下での自動採点プログラムの精度測定

本項では実際の採点作業計画を念頭にした場合に想定される作業工数について測定、推定します。

3.3.3.1 必要な先行採点件数の推定

教師データの件数を増やすほど自動採点の精度が高くなることが期待できる一方、先行採点のための採点期間・工数が必要となります。

このため、本項では教師データの件数に応じた自動採点の精度を測定することで、先行採点に工数を投入することの有効性について推定を行います。

教師データ件数に応じた、自動採点プログラムによる採点精度の測定

本項では、教師データの件数を、200 件（半日以内で教師データを作成し、分析のスケジュールを１日早めることを想定した件数）、500 件（項番 3.3 で示した内容）、1,000 件（１問あたりの上位採点者を２倍とするなど、先行採点にリソースを投入した場合を想定した件数）で変化させた場合の、それぞれ５回の測定結果を示します。

なお、小学校算数では、項番 3.3 で述べたように正しい選択肢を選び、内容解釈の精度測定対象として妥当な解答が 637 件であったことから、教師データ 1,000 件による結果の記載はありません。

プログラム A_小学校国語

類型一致率	教師データ 200 件	教師データ 500 件	教師データ 1,000 件
平均値	71.73%	75.70%	77.35%
最大値	74.72%	76.57%	77.74%
最小値	69.78%	74.69%	77.01%

プログラム A_小学校算数

類型一致率	教師データ 200 件	教師データ 500 件
平均値	66.00%	71.82%
最大値	69.79%	76.64%
最小値	62.93%	69.34%

プログラム B_小学校国語

類型一致率	教師データ 200 件	教師データ 500 件	教師データ 1,000 件
平均値	52.86%	67.88%	79.31%
最大値	55.28%	73.58%	81.28%
最小値	48.50%	59.23%	77.33%

プログラム B_小学校算数

類型一致率	教師データ 200 件	教師データ 500 件
平均値	43.94%	53.43%
最大値	50.11%	62.04%
最小値	40.64%	48.18%

特に生成 AI の原理を用いたプログラムでは、教師データの件数を減少させると言語モデルに対するファインチューニングが十分に行われず、文章と分類すべき値との結びつきなしに分類が行われてしまうことが推測されます。このため教師データ 200 件とした際のプログラム B では、特定の値に偏った出力が多発し、精度についても非常に低いと言わざるをえない結果となりました。

プログラム A でも分類はなされるものの精度が大幅に落ち込んでおり、教師データ作成の時間短縮以上に品質向上のための工数が増加するため、結果として効率化につながらないことが強く疑われます。

一方で、教師データを 1,000 件に増加させると両プログラムとも精度が向上し、特にプログラム B ではプログラム A を上回る測定結果となりました。また、プログラム B で生じていた、生成 AI 特有の特定の値に偏った出力も発生しませんでした。

精度１%の差は悉皆調査の規模では約１万件の差に相当することから、上位採点者の採用・研修体制を整えることで、早期に可能な限り多くの種類の教師データを収集することが望まれると考えられます。逆に、教師データの件数を減らすことは、悉皆調査の規模では分類精度が低下することによる影響が大きく、技術的裏付けや採点工程の事情がない限りは、早期化にはつながらないと推測できます。

3.3.4 中学校理科に対して、今回の精度測定結果を当てはめる妥当性の考察

今回選定した問題は、選択肢等の形式的に判断される条件１つと、内容に関して判断する２つ～４つの条件を正しく判断することで分類されます。

一方で、令和 4 年度本調査中学校理科で出題された 5 問について、類型分類のために必要な条件を整理すると次の表のようになります。

設問番号	条件①概略	条件②概略	条件③概略
4(1)	資料にある生物の生活場所を比較している	資料にある生物の移動手段を比較している	２つの生物に具体的にふれて解答している
5(3)	実験で必要な測定値の刻み幅を数値で示している	実験で必要な測定範囲を数値で示している	測定する際の具体的な値を示している
8(1)	実験により生物の動き方が変化しなかったことを記述している	資料にある生物が視覚情報を処理していないことを記述している	実験で調べていない観点の考察をしていない
8(2)	実験の条件制御が不十分だったことを指摘している	実験に必要な操作が不十分であることを指摘している	所定の操作手順に従わず実験したことを指摘している
8(3)	昆虫が持つ体の特徴を指摘している	未知の節足動物が持つ体の特徴を指摘している	２種類の生物の特徴を比較している

これをふまえ、本報告書で選定した問題と、中学校理科で出題される問題との間にある、自動採点を行う上での特徴の差については、以下の 2 点を考慮する必要があります。

観点	考慮要因
中学校理科の問題特性	・正答の条件を満たすために記述しなければならない解答の方向性が限定されており、「自分の考え」のように抽象的な条件で無いという点で、国語の記述式問題よりも採点が容易と考えられる。 ・原則として言葉による解答を求め、数式による解答を要しない点で、算数の記述式問題より採点が容易と考えられる。
本報告書における選定問題の特性	・解答内容が多様で人手による採点難易度が高い問題を優先して選定した経緯があることから、教科を問わずほとんどの問題は、今回の問題よりも採点が容易と考えられる。

以上より、中学校理科の解答群を対象とした自動採点においては、正答の条件の数が異なることを加味してもなお、今回の測定結果に基づいた推計と少なくとも同程度の精度による自動採点が可能と見込まれることから、項番 4 以降で述べる自動採点プログラムを導入した際の効率化の程度の推計にあたっては、本項番 3.3 の中で測定した結果をもとに推計を行います。

3.4 実施した自動採点の工数測定

3.4.1 システム運用の工数の測定

本項では、自動採点プログラムによる解答分類で生じた、処理時間に関する測定結果を示します。

なお、この結果は項番 3.1 で示すプログラムをもとにしており、細かな工程や処理時間等は、受託事業者等が保有するそれぞれのプログラムにより異なることに留意する必要があります。

悉皆調査時に求められると考えられる処理時間に関する仕様については、項番４にて示します。

3.4.1.1　　自動採点プログラムの搭載、システム構築に必要な作業工数

本事業で実行した自動採点プログラムのシステム構築・搭載については、採点に必要な解答データを受領するまでに準備が完了しており、システム構築の部分で、採点工程に影響するタイムラグは発生しませんでした。

加えて、採点データ受領から自動採点プログラムの実行までに、次の表に示すデータ処理を行いましたが、これらの工程は１営業日以内で完了しています。

工程	概要
ファイル内容の確認	受領データのファイル名、列名、各列に対応する入力情報の内容や入力ルール、取りうる値の範囲等フォーマットの確認
採点結果の変換	事業者の採点システムに対応した入力内容を、採点マニュアルに照らして、正答の条件別の正誤情報へ変換
自動採点プログラムへの登録	自動採点プログラムに必要な情報の抽出、システムへのアップロード等

悉皆調査についても、自動採点プログラムを既に有している事業者については、提案される自動採点プログラムを実行するためのシステム構築は調査実施日までになされると推測できますが、仮に新規のプログラムを開発・搭載する際には、調査実施日までに開発および動作テストを完了するための方針が示される必要があると考えられます。

加えて、特に MEXCBT 等から出力されるデータを速やかに自動採点プログラムで実行可能な形に変換する体制について考慮することが求められると考えられます。

3.4.1.2　　解答分類に必要な処理工数の測定

本事業で用いる２つのプログラムを用いた自動採点では、概ね下に示す工程を経て、未知の文章に対して類型や正誤が出力されます。

	プログラム A （機械学習手法を用いた分類）	プログラム B （生成 AI の原理を用いた分類）
工程０	児童生徒の解答データの取得、自動採点プログラムで採点するためのデータ作成	
工程１	人手による先行採点に基づき、解答文と出力する値の対応一覧を作成	
工程２	解答文のベクトル化	先行採点の結果に基づく言語モデルのファインチューニング
工程３	先行採点の結果を教師データとして分類モデルを作成し、採点されていない状態の検証データを分類	ファインチューニング後の言語モデルを用い、残りのデータを分類
工程４	自動採点の結果を、人手による採点システムに連携	

このうち、工程１～工程３に関する測定結果、推計内容について下に示します。

工程１：先行採点に必要な工数の推定

項番 3.3.3 で示唆されたように、先行採点を行うほど教師データが多数得られ、自動採点プログラムの精度向上につながりますが、一方で、先行採点の件数を増やすには先行採点のための工数と期間を確保する必要があるため、通常の人手による採点の期間を圧迫することが懸念されます。

解答の分類に必要な処理時間や、データ処理・人手による採点システムへの連携に必要な工数を考慮すると、従来どおりの日程で一般採点者による採点を開始するためには、先行採点の期間は１～２日程度の期間にとどめ、自動採点プログラム全体としての工数と期間を抑えることが必要と考えられます。

このため、項番 3.3 で実施した自動採点プログラムの精度測定にあたっては、解答１件あたり採点に１分かかるような難度の高い文章でも、上位採点者１名が１日（約８時間）程度で１回採点を行って教師データを作成できる件数として、教師データの件数を 500 件と設定し、測定を行っています。

また、本報告書の自動採点精度の測定・検証では、上位採点者が 1 回のみ採点した結果を教師データとして用いましたが、効率化につながることが充分確認できる程度の精度で自動採点が行われています。悉皆調査においても、上位採点者が採点する等、採点結果に対し十分に高い品質が期待できる場合には、２回採点や最終的な品質確認を要さずに教師データとして自動採点プログラムに用いることで、より効率的な採点進行が可能と考えられます。

工程２～３：プログラム A（機械学習による分類）に必要な処理工数の測定

・工程２：解答文のベクトル化に必要な時間の測定・推計

解答をベクトル化するなどして、児童生徒の解答を機械学習に適したフォーマットに変換する際には、解答の１行ずつを埋め込みモデル等を用いて変換していくため、一般にはその行数に概ね比例した処理時間が必要と推測されます。これに加えて、処理件数が多くなるほど処理すべきファイルの数やサイズが大きくなるため、システムへの負荷や入出力部分の処理時間がより増加する可能性に留意する必要があります。

測定結果もふまえ、悉皆調査で想定される、1 問あたり 100 万件規模・複数問題の解答データに対する変換処理では、処理工数が採点工程全体を圧迫しないように、システムの並列化などにより、処理時間の短縮が必要となることが考えられます。

・工程３：分類モデルの作成、解答の分類に必要な時間の測定・推計

分類モデルの作成から自動採点処理を完了するまでの時間の差は、原則としては教師データの件数と分類に必要な条件の数に依存すると考えられます。

同じ教師データ件数でも教科間で 1 条件あたりの処理時間が異なった要因としては、データの入出力の時間や、プログラムを動作させているシステム全体の負荷状況、各教師データの内容をもとにした、分類モデル計算の収束の速さの違いなどが推測されます。

なお、工程２（解答のベクトル化）と同様に、本工程でも悉皆調査の規模でのデータ数を処理する際には、ファイルの大きさに対応するシステムの構築や、システムの並列化などによる時間短縮が望まれると考えられます。

工程２～３：プログラム B（生成 AI を用いた分類）に必要な処理工数の測定

・工程２：言語モデルのファインチューニングに必要な工数の測定・推計

ファインチューニングに必要な時間は、プログラム A の分類モデル作成と同様、原則としては教師データの件数と分類に必要な条件の数に依存すると考えられます。

プログラム A と比べて、同じ教師データ件数のとき教科間でファインチューニング時間の差が小さくなった理由としては、プログラム A に比べ、処理で生じる負荷に対し余裕のあるハードウェア構成であった点や、実行するプラットフォームそのものが、リモートデスクトップと専用サーバーで異なったことに由来すると推測されます。

・工程３：分類結果の出力に必要な時間の測定・推計

生成 AI を用いた分類の場合、入力した文章をもとに、生成した文章を分類結果として扱うことで処理を行うため、処理時間は原則として入出力の行数に比例すると考えられ、測定結果からも、１行１条件あたりの処理秒数は概ね等しく、採点に必要な時間は解答件数と正答の条件の数に概ね比例すると考えられます。

3.4.2 悉皆調査時の、自動採点プログラムの運用・処理に関する時間の推計

ここまでの測定結果を整理すると、それぞれの工程の処理時間は、次の表に示す要因によって主に増減すると推測されます。

要因・変数	プログラム A ※機械学習手法による分類	プログラム B ※生成 AI の原理を用いた分類
教師データの数	・先行採点の工数 ・分類モデルの作成時間＋分類結果の出力時間	・先行採点の工数 ・ファインチューニングの処理時間
受験人数	・ベクトル化の処理時間	・分類結果の出力時間
正答の条件の数	・分類モデルの作成時間＋分類結果の出力時間	・ファインチューニングの処理時間 ・分類結果の出力時間
出題問題数	・上に示す全ての工程の処理時間	

悉皆調査時の運用・処理の際には、これらの要素を念頭に処理時間を考慮する必要がありますが、先行採点を除く工程については、多くの場合システム構成の増強と並列化によって解決可能であることから、原則としては各工程の処理時間については、１日から最大２日程度の単位で考えることができると推測されます。

項番３のなかで示した測定・推計をもとに、自動採点プログラムの運用・処理に要する工程と、想定される期間の概算を下に記載します。

時期	項番 3.4.1.1 との対応	工程の概略	確保期間の概算
調査実施日～	工程０	児童生徒の解答データの取得、自動採点プログラムで採点するためのデータ成形	１日～２日程度 ※調査実施日ごとに適宜対応
２日後～３日後	工程１	人手による先行採点、教師データの作成	１日程度
２日後～５日後	工程２	解答文のベクトル化、ファインチューニング	１日程度
	工程３	解答データの分類、自動採点処理	１日～２日程度
～１週間後	工程４	自動採点の結果を、人手による採点を行うシステムへ連携	２日程度

これにより、従来人手による採点を行う際にも確保されている、概ね 1 週間程度の先行採点期間のなかで、自動採点プログラムによる解答データの分類を完了することが期待されます。先行採点期間の中で自動採点を完了できれば、これまでの人手による採点では未採点の状態から一般採点者による作業を開始するところを、1 回以上採点が完了した状態から開始するように扱うことができます。

3.4.3 上位採点者による類型確定に必要な工数の測定

人手による採点と自動採点の結果が一致しなかった場合、上位採点者による最終判断が必要となります。本項では人手による判断との一致率が最も低かった回の採点結果を抽出して、判断が一致しなかった解答に対して最終判断を行う際の工数について測定します。

なお、作業プロセスは過年度の人手による採点で行われている内容を踏まえ、不一致が生じた解答における人手による採点結果とプログラムによる採点結果を、どちらが誰によるものかは明かさずに上位採点者に対して渡し、改めて上位採点者自身で判断するよう指示を行いました。

プログラム A で測定したもののうち、最も類型一致の精度が低かった回で生じた不一致に対する、上位採点者による類型確定の測定結果を、次の表に示します。

教科	件数	作業時間	1 件あたり作業時間	1 万件あたり作業時間 ※不一致 1%に相当	プログラム A_教師データ 500 件の不一致率の場合
小学校国語	928 件	230 分	14.87 秒	約 41 人/時	約 996 人/時
小学校算数	42 件	20 分	28.57 秒	約 80 人/時	約 2,254 人/時

上位採点者に対するインタビューからは、自動採点と人間との間の判断不一致は、記述内容の解釈よりも、採点基準の理解に由来する物が多く、自動採点後の解答結果に対する品質向上も、人間のみで行う場合と同様、採点マニュアルの適切な運用によって実施できることがうかがえます。

加えて、採点マニュアルの記述で解決できる内容が多くを占めることから、採点マニュアルの内容と自動採点の結果を照らし合わせることで、事前に自動採点プログラムが判断を誤りやすい箇所を特定して、品質向上の観点に組み込む手法が考えられます。

採点秒数としてはこのあと項番４で示す PBT に対するものより短い結果となりましたが、これは CBT 特有の要因として、手書き答案ではなく CBT のテキストデータを読むために目視が容易なことや、csv ファイルを操作することによる採点のため、単語検索やソート等を併用することで、解答画像を目視確認する PBT の採点システムよりも、採点時の負荷が小さかったことに由来すると推測されます。

ただし、悉皆調査の規模では、合計としては上位採点者が行う品質向上作業として相応の工数を割く必要があると見積もられます。加えて、自動採点の分類精度が人手による精度よりも高くない現状では、分類誤り等によって上位採点者が最終判断を下す必要がある解答の割合が、人間のみの採点プロセスよりも増加する可能性も考えられることから、上位採点者の確保や、採点者全体への研修体制といった、人間による採点における品質担保の仕組みは、引き続き重要な観点であると推測されます。

3.4.4 検収完了までに必要な工数の測定

検収完了までに必要な工数の測定として、項番 3.4.3 で作業を行った者とは異なる上位採点者によって、採点結果を全件確認しました。

検収手順については全国学調で行われる内容を踏まえ、下のとおり行いました。

手順	概要
手順１	項番 3.4.3 の採点結果を別の上位採点者によって確認し、正誤の分類誤りがないかを確認する。
手順２	正誤が一致しない採点結果が確認された場合は、不合格となった類型のみを項番 3.4.3 で作業を行った採点者に伝え、どの解答で誤分類が生じたか不明な状態で、再度採点結果に対しての確認を行わせ、完了後再度検収を行う。

完了までに必要な検収の回数と、再度の品質向上作業が行われた割合は、次の表のとおりです。

教科	不合格回数	不合格類型数	述べ品質再確認対象割合
小学校国語	１回	２類型	27.16%
小学校算数	１回	１類型	26.19%

記述式設問をはじめとした解答内容を判断する採点では、機械による判断でも人間による判断でも誤る可能性があるため、その採点内容によっては再検収が行われる可能性があり、再検収となった対象範囲に対し、追加の品質向上作業が発生することを念頭に置いて、採点計画を組み立てることが望まれます。

3.5 その他経年変化調査問題の採点精度

本項では、その他経年変化調査問題に対して、自動採点プログラムを使用した際の採点精度の一覧を示します。

なお、採点方法の説明は、下の表のとおりです。

項目	測定内容
使用するプログラム	教科・問題にかかわらず、全問題でプログラム A（機械学習による分類）を使用した場合の精度を測った。
教師データの件数	前項までの内容から、人間による教師データ作成が可能な範囲内で最も精度が期待される 1,000 件を教師データとした。 なお、項番 3.3.3 で選定した算数の問題（表では“算数 X”と表記）については、前述のとおりデータ件数の不足から 500 件を教師データとしている。

以降の表で示す測定結果の項目と示す値の内容は、下の表のとおりです。

表の項目	記載内容
問題記号	本報告書内でまとめるため、実際の問題管理記号とは別に便宜上割り振ったアルファベット。 なお、国語と算数の一番上の行で示す”問題 X”は、項番 3.3.3 で選定した問題である。
正答の条件数	各設問で設定されている、内容の判断を要する正答の条件の数。 解説資料等で記載される正答の条件の数に準拠している。 また、「正しい選択肢を選ぶ」や「適切な字数で書く」など、形式的・関数的な手法のみで確定できる正答の条件は除いている。
類型数	各設問で設定されている、解答類型の数。 なお、無解答や選択肢を誤った場合に分類される類型のように、形式的・関数的な手法で一意に定まる類型は除いている。
類型一致率（平均）	人間が採点した際の解答類型と、自動採点プログラムにより出力された解答類型が一致した割合の平均
正誤一致率（平均）	人間が採点した際の正誤と、自動採点プログラムにより出力された正誤が一致した割合の平均

3.5.1 各教科の採点精度の一覧と概況

3.5.1.1　　小学校国語出題問題に対する採点精度

問題記号	正答の条件数	類型数	類型一致率（平均）	正誤一致率（平均）
国語 X	2	4	77.35%	89.42%
国語 A	1	2	90.67%	90.67%
国語 B	2	4	93.90%	95.85%
国語 C	2	4	85.09%	89.47%
国語 D	2	4	84.61%	92.85%
国語 E	2	4	93.63%	97.69%

国語については類型一致率・正誤一致率ともに比較的高い結果が得られており、文章の内容に対し機械学習等を用いて分類する自動採点プログラムの機能は概ねはたらいていたと考えられます。

項番 3.3.2 で選定した問題のように長文の記述を求める問題では一致率が低下するものの、項番 4 で後述するように、この程度の類型一致率であっても、自動採点プログラムにより採点工数の効率化には寄与すると推計されます。

3.5.1.2　　小学校算数出題問題に対する採点精度

問題記号	正答の条件数	類型数	類型一致率（平均）	正誤一致率（平均）
算数 X	4	5	71.82%	85.99%
算数 A	3	6	74.02%	86.92%
算数 B	3	7	71.69%	91.83%
算数 C	4	6	83.90%	87.23%
算数 D	2	4	86.49%	89.05%
算数 E	7	9	86.59%	95.56%
算数 F	3	4	87.85%	90.28%
算数 G	4	6	72.65%	95.20%
算数 H	5	9	80.72%	92.70%
算数 I	2	6	77.30%	88.88%
算数 J	4	5	73.39%	89.61%
算数 K	12	8	66.66%	84.49%
算数 L	12	8	68.25%	88.96%

算数については、類型一致率が国語に比べ低下した一方で、正誤一致率については国語と遜色ない精度が得られています。これは、正答の条件が他教科に比べて特に多くなった傾向から、正誤を分ける観点までは正しく分類できても、正誤から細かく類型を分けるための観点までを含めて全ての条件までは正しく分類できなかったケースが増加したためと推測されます。

このため、正答の条件や類型が多数存在するような問題に対しては、正誤の分類に直結する観点を自動採点で実施し、そこからの詳細な類型への分類を採点者で行うような手法が考えられます。

3.5.1.3　中学校英語出題問題に対する採点精度

問題記号	正答の条件数	類型数	類型一致率（平均）	正誤一致率（平均）
英語 A	3	4	70.22%	90.19%
英語 B	3	4	55.17%	77.35%

英語については、短い文章で答える問題（問題記号：英語 A）でも類型一致率が低く、長文記述を要求する問題（問題記号：英語 B）では正誤一致率まで拡大しても非常に低い精度となりました。

特に英語に対する採点に際しては、文法の正しさを判断する観点から 1 文字の違いが類型や正誤に直結する傾向が他教科と比べて非常に強く、そのような誤りを適切に検知できなかったことが要因として考えられます。

また、長文記述を要求する問題では、他教科と比べて正答と認められる記述内容の判断が非常に広く、児童生徒の解答としても正答誤答ともに非常に多様な解答が得られたことから、教師データでカバーしきれない範囲の解答内容を分類する必要が生じ、文法など 1 文字の違いに由来する類型検知の困難さと相まって、採点精度が非常に低下してしまったことが考えられます。

3.5.2 不一致を生じた解答傾向の質的分析

本項では、分析対象とした問題において経年変化調査実施時に人間が与えた正しい採点結果と自動採点による採点結果との間に不一致が生じた解答や、採点精度が他と比べて低くなった問題の出題内容について確認し、自動採点を実施した際に採点精度に影響を与える要因について考察します。

なお、本事業では非公開問題を使用していることから、本項の内容については個々の問題内容や具体的な正答の条件等が明らかとならない範囲で、過去本調査で出題された問題も参考にした傾向のみを示します。

観点	留意点
正答の条件・類型の数	当然のことながら、判断すべき正答の条件や分けるべき類型の数が多くなるほど、全てを正しく判断できる確率は低下し、併せて採点精度も低下する。 一方で、類型が異なる採点結果でも結果的に正誤は一致するケースにより、正誤一致率の低下割合は類型一致率の低下割合よりも抑えられる。 また、反応率が非常に低い類型や正答の条件が複数存在するような場合、類型や正答の条件の増加が採点精度の低下に直接つながるとは限らない。
形式的に一意に定まる採点基準	特に今回用いたような自動採点プログラムでは、単語や文章の意味内容を数値に変換して機械学習などを行うため、意味内容に依らない正答の条件に対しては、分類を誤る可能性がより高くなる。 このような採点基準を持つ問題に対しては、項番 3.3.3 の中でも行ったように、字数のカウントや語句検索など、関数的な処理を言語モデルによる内容判断と組み合わせることで、より精度の高い採点結果を得られると考えられる。
多様な解答内容	１つの正答の条件の中で多様な解答内容が認められる場合、異なる意味内容の文章に対し同じ判断を行うよう採点基準で求められたり、教師データ作成時に想定しなかった解答内容を採点したりするなどで、分類モデルに対する採点精度が低下しやすくなると推測される。 このような解答傾向を持つ問題に対しては、あらかじめ記述内容で解答群をクラスタリングするなど、特定の内容に集中した教師データの作成や、１つの正答の条件を内容別に更に細かく分割して自動採点を行うなどで、より精度が向上する可能性がある。

3.6 本事業テーマ B 実施問題に対する分析

本項では、本事業テーマ B（”CBT での調査実施等に関する試行・検証”。以降「テーマ B」と表記）のなかで出題された問題に対する分析結果について報告します。本年度テーマ B における作問の都合上、テーマ B の試行・検証のなかでは記述式問題が出題されなかったため、本項では短答式問題を対象とした場合の、CBT 実施と自動採点による効率化の程度を推計します。

しかしながら、計算問題や単語の穴埋め問題のように、各類型に当てはまる解答が 1 つか少数しかない場合には、表記ゆれなどを考慮しても選択式の問題と同様に、採点の際には CBT で集められた解答データに対し各類型に当てはまる文字列を単純に検索して類型を与えることで、十分な効率化が達成できることは明らかです。

このため、本項では各類型での許容表現が列挙できない程度に多数ある問題として、中学校国語１問、中学校英語 5 問を対象として、短答式問題においてどの程度の効率化がなされうるか推計しました。

3.6.1テーマ B で出題された短答式問題に対する推計

推計にあたり、複数の受検者がそれぞれ同一の文字列で解答しているものが、正答誤答を問わず多数存在することに着目しました。

特に知識・技能を問うような形式の短答式問題については、許容される表現が多数あったとしても、実際に児童生徒が解答する文字列としては設問に沿ってある一定の範囲に集まると推定され、それら典型的な解答文字列に対して一括で採点を行うことができれば、解答 1 件ごとを目視採点するのに比べ、大幅な効率化が達成できると推測できます。

対象とした短答式問題について、解答文字列の種類数と特に頻出する解答が全体に占める割合をまとめたものが下の表です。

設問記号	受検者数に対する 解答種類数の割合	頻度上位 10 件の解答文字列と 無解答が、全受検者に占める割合
国語 A	29.60%	64.61%
英語 A	23.86%	65.28%
英語 B	32.83%	45.35%
英語 C	22.52%	63.62%

設問記号	受検者数に対する 解答種類数の割合	頻度上位 10 件の解答文字列と 無解答が、全受検者に占める割合
英語 D	25.98%	65.04%
英語 E	36.85%	45.35%

左の列の値について、例えば 1000 人の受検者に対して、異なる文字列として区別される解答が約 225 種類～約 368 種類であったことを表します。このため、重複する解答に対して一括で採点を行える仕組みを有する採点プロセスであれば、従来 1000 件の解答に対して行っていた採点判断を 300 件前後の判断で完了できることを表します。

また右の列の値について、典型的な解答文字列 10 件と無解答に対して機械的に類型を与える機能があれば、それによって約 45.35%～約 65.28%の解答群に対して採点を終えることができることを表します。また、受験者数が多くなればなるほど、典型的な解答文字列に対して一括で採点を行う機能の効果は相対的に大きくなります。

頻出解答の傾向や割合は、解答者群が持つ能力の分布を同一と仮定すれば、十分に受検人数が多いときここに示したものと概ね同程度の割合であると推測することができます。100 万人規模に対する採点であっても、許容される表現が多数あっても実際の児童生徒の解答範囲がある程度に収まると想定される問題に対しては、出現頻度が高い解答に対して一括で採点を行う仕組みの導入により、全ての解答を逐一目視採点する場合に比べ、ここに示した程度の作業工数の圧縮が期待できます。

一方で、このように短答式問題に対し効率的な採点プロセスを導入できた場合であっても、100 万人規模の調査実施では、引き続き多数の解答を人間が判断する必要がある点には留意が必要です。

受験者数に対する解答種類数の割合を 30%と仮定した場合、100 万人に対しては 30 万種類の解答に対して判断を下す必要があります。このような場合、頻度の高い解答に対しては単語検索などを用いて一括で採点しながら、少数解答に対しては意味内容による自動採点を組み合わせるなど、特定の採点手法にこだわらない、問題の特性に応じた柔軟な採点手法の運用が望まれます。

4 全国学力・学習状況調査（悉皆調査）での自動採点の活用可能性

本項では、全国学調経年変化調査の中で実際に収集されたデータと照らして、自動採点プログラムの採用によってどの程度の効率化が期待できるかを示します。

4.1 悉皆調査の採点を人手のみで行った場合の工数推計

本項では、項番 2 の中でとりあげた、従来人手のみで行われている採点工程において、どの程度の工数が必要と考えられるかを確認します。

4.1.1 人手による採点に関する条件の設定

実測値をもとにした、人手による採点に関する諸条件の数値設定

現在人手による採点では、パソコンの画面上に表示される解答画像を目視し、正答の条件の適否や類型などを入力していく手法が一般的に用いられています。CBT による採点においても、表記される情報が解答用紙とテキスト情報の違いはあるものの、品質向上作業など、多くの作業は従来の採点方法と同様に行われることが想定されます。

このため、工数の推計のために必要な PBT 実施時の採点工程に関する実測値について、全国学調の小学校事業者より情報の提供を受け、 1 件あたりの採点完了までに必要な秒数について推計を行いました。工数推計に際して提供を受けた情報を下に示します。

なお、この値は経年変化調査の採点を担当した、上位採点者以上の立場にある者を対象としたものであり、悉皆調査の採点時に多数を占めると想定される一般採点者の場合には、より採点に時間がかかり、かつ誤分類の割合が高くなる点には留意する必要があります。

観点	小学校国語の実測値
1 件あたりの採点秒数	29.27 秒 180 分/ 369 枚
1 回目採点で判断を保留した解答の割合	16.99% 729 件/4291 件
採点の判断に対して誤分類が生じた割合	4.46% 159 件/3562 件

※小数第 3 位で四捨五入

人手による採点に関する諸条件の仮定

人手による採点に関して、実測できない条件については、過年度の実施状態を参考に、次の表の通り仮定した推計を行います。

観点	設定・仮定
２名の採点者の判断が一致しない解答・一般採点者が保留した解答に対する採点	保留された解答や不一致となった解答に対しては上位採点者による採点が行われますが、これらも通常の採点と同様の速度・品質で判断されると仮定します。
品質確認の時間	特定の類型に限った品質確認など、通常の採点と比べて目視難易度が下がるため、品質確認の時間は通常の採点の 0.5 倍と仮定します。
品質確認の精度	品質確認の作業でも、通常の採点と同様の割合で誤分類（正しい採点結果を誤りとして検出／誤った採点結果を見逃す）が発生すると仮定します。
解答の最終判断	品質確認時点で疑義が生じた解答は、上位採点者の合議等によって判断されるため、最終判断で分類された解答の採点精度は100%と仮定します。 また、合議等による判断を行うことから、最終判断に必要な採点時間は通常の２倍（２名分）として計算します。

4.1.2 人手のみによる採点を行った場合の工数推計

項番 4.1.1 で仮定した小学校国語の人手による採点能力（約 29 秒で採点を行い、約 4.5%で誤分類が生じる）を例として、解答 1 件あたりの人手による採点工数を推計すると、次の表のように表すことができます。

工程	1 件あたり秒数	正しく採点された割合	誤分類の混入割合	未分類状態の解答の割合
1 回目採点	29.27 秒	–	–	–
2 回目採点	29.27 秒	75.76%	0.17%	24.07%
保留・不一致解答の判断	7.05 秒	98.76%	1.24%	0%
品質確認①	14.64 秒	94.36%	0.06%	5.59%
未分類解答の最終判断①	3.27 秒	99.945%	0.055%	0%
合計の採点秒数	83.50 秒	–	–	–

※小数第 3 位で四捨五入、最終的な精度のみ小数第 4 位で四捨五入、以降同様

上の表で行われている試算について、人間では判断に迷う答案の割合が全体の 16.99%、採点したときに判断を誤る割合が 4.46%という条件で 2 回採点を行うと、1 回目の時点で判断が保留された答案と、2 回の採点結果が一致しない解答の合計が約 24.07%となり、それに対して再度の採点が必要となります。

品質確認の際には、誤分類の 95.54%を正しく検知するのと同時に、正しい分類を 4.46%の割合で誤って検知してしまうため、結果として上位採点者が最終判断を行わなければならない解答の合計は 5.59%になる、という流れで総工数が求められています。

人手による採点では、2 回採点に加えて 1 回の品質確認を行うことにより、誤分類率が 0.05%程度になったと見積もられ、そこまでにかかった採点工数を 1 件あたりに均すと、約 83.50 秒必要であったと推算されます。

4.2 悉皆調査の採点で自動採点を併用した場合の工数推計

本項では、測定結果の値に基づき、自動採点プログラムを使用した場合の工数削減の程度について示します。

4.2.1　1 種類の自動採点プログラムを併用した場合の採点の工数推計

機械学習により、74.69%（項番 3.3　表②-2-1 で示した、ルールベースの処理を加えた上での、類型一致率の最小値）の精度で全ての解答に対して 1 回自動採点を行い、その後項番 4.1.2 と同様に 2 回目以降を人手による採点で行った場合の採点工数を推計すると、次の表のように表すことができます。

工程	1 件あたり秒数	正しく採点された割合	誤分類の混入割合	未分類状態の解答の割合
自動採点	−	−	−	−
人手による 1 回採点	29.27 秒	59.23%	0.94%	39.83%
保留・不一致解答の判断	11.66 秒	97.28%	2.72%	0%
品質確認①	14.64 秒	92.95%	0.12%	6.93%
未分類解答の最終判断①	4.06 秒	99.88%	0.12%	0%
品質確認② ※全体の 6 割を確認	8.78 秒	97.208%	0.052%	2.74%
未分類解答の最終判断②	1.61 秒	99.948%	0.052%	0%
合計の採点秒数	70.02 秒	−	−	−

自動採点プログラムを 1 回採点に置き換えた場合には、自動採点に加えて人手による 1 回採点を行い、それに対し 1.6 回分の品質確認を行うことにより、人手のみで実施した場合と同等の品質が達成できると見込まれます。

2 回の採点結果が一致しない解答の割合や、混入している誤分類の割合が増加するため、上位採点者の工数は増大しますが、一般採点者の採点回数が減少することにより、全体としては人手による採点と比較して、約 16.14%の秒数減少となることが見積もられました。

この推計は上位採点者の値をもとにしていることから、自動採点プログラムとの品質の差がより小さい一般採点者による採点工程では、自動採点に置き換えることによる減少割合は、これより大きくなることが期待できます。

4.2.2　悉皆調査実施時の工数推計

項番４で推計の対象とした問題は、項番 3.3.1 で示した通り、人手による採点と、自動採点による分類がともに困難と考えられる問題を選定しています。項番 3.3.4 で示したように、中学校理科で出題される各問題における採点の難易度はこれよりは低いと考えられることから、今回測定・推計した内容と概ね同等以上の割合で、採点秒数の減少につながると推定され、CBT 導入後の採点作業の工程については、項番４で示された割合をもって効率化がなされるであろうと推定したスケジュール設定が可能と考えられます。

ただし、採点作業全体でどの程度の工数、作業人数を要するかは、出題される記述式問題の数や各問題の解答記述量、正答の条件の数に強く依存するため、採点作業工数の短期間化については、出題する問題の内容にも留意が必要です。

また、自動採点プログラムの運用に関する工数や、CBT の解答データを人手による採点工程に連携する部分などは、従来の採点作業では発生しない部分であるため、これらに由来して想定よりも作業工数が増大する可能性も考慮して、採点計画等を設計する必要があると考えられます。

4.3 令和７年度調査で自動採点を活用する場合のスケジュール（素案）

特に中学校で調査を実施する３教科について、人手による採点と自動採点プログラムによる採点それぞれで想定される工程、処理について次の表に示します。

<table>
<tr><th>日程</th><th>人手による採点の工程</th><th>自動採点プログラム等の工程</th></tr>
<tr><td>～4月 14 日</td><td>・採点基準の作成
・採点マニュアルの作成
・採点者の採用</td><td>・システムの構築
・プログラムの搭載、想定解答例等のテストデータを用いた動作確認</td></tr>
<tr><td>4月 14 日</td><td>・理科調査実施開始</td><td>・MEXCBT からのデータ収集開始</td></tr>
<tr><td>～4月 17 日
※調査実施日</td><td>・国語、数学の調査実施（4 月 17 日）
・主に教科責任者による、理科の先行採点開始</td><td rowspan="2">・理科の自動採点を優先実施
・他教科の自動採点実施
※自動採点の各工程は項番 3.4 で記載</td></tr>
<tr><td>4月 17 日～１週後</td><td>・国語、数学の先行採点開始
・理科の自動採点結果確認
・人手による採点システムとのデータ連携</td></tr>
<tr><td>１週後～３週後</td><td>・一般採点者による採点開始
・自動採点結果の内容確認、品質向上方針決定</td><td rowspan="3">・後日実施解答等の自動採点実施</td></tr>
<tr><td>２週後～４週後</td><td>・理科の品質確認、品質向上作業開始</td></tr>
<tr><td>３週後～４週後</td><td>・理科の品質向上完了、検収
・国語、数学の品質向上作業開始</td></tr>
<tr><td>４週後～６週後</td><td>・国語、数学の品質向上完了、検収</td><td>・理科の採点結果を用いた IRT による分析の開始</td></tr>
<tr><td>６週後～</td><td colspan="2">・全問題採点完了、集計業務の開始</td></tr>
</table>

4.4 自動採点プログラムの仕様として求められる観点

本項では、ここまでに示した測定結果、推定結果をもとに、特に令和７年度中学校理科における自動採点プログラムを利用した採点に関し、提案を受けることが望ましい観点について整理します。

ただし、本項で取り上げられる内容は採点工程の全体最適化に関する観点となることから、各項目の内容は相互に関連し、最終的には総合的な判断が望まれる点に留意する必要があります。

4.4.1 自動採点プログラムの構築に関する要件

4.4.1.1 プログラム・アルゴリズムの要件

今回の測定結果では、機械学習や生成 AI 技術を使わず、従来から存在するルールベースの分類（例：字数の計測、単語検索）などを用いた方がより高い精度が得られる場合が考えられました。また、処理時間の観点などからも、効率的に完了できる自動採点の手法は１つに定まらないと考えられます。

このため、必須要件として特定の分類手法・アルゴリズム等を求めるのは好ましくなく、今後仮に具体的な手法を仕様書等に記載する場合でも、イメージ共有のための例示等にとどめるのが望ましいと考えられます。

ただし、MEXCBT が利用されるなどの理由によって、ファイル形式をはじめ自動採点を行う上でのフォーマットが一部定まっている場合には、例えば MEXCBT で収集されるデータを処理するためのプログラムを事前に仕様として要求するなどにより、自動採点プログラムと解答データの連携の部分で余分な工数が生じる可能性を減らせると考えられます。

4.4.1.2 セキュリティ・ネットワークの要件

まず、人手による採点は引き続き行われることから、令和６年度までに示されていた採点工程に対するセキュリティ要件については、引き続き遵守されることが強く望まれます。

その上で、本事業の仕様としては自動採点プログラムの構築について次の表に示す要件が示されており、本報告書で取り上げたプログラムはこれを遵守して構築され、試行・検証が行われました。

このため、悉皆調査についてもこれに準じた要件のもと、自動採点プログラムのシステム構築が望まれると考えられます。

項目	本事業仕様書の記載
オープンソース等の商用利用	第三者により作成されたオープンソース等の活用にあたっては、許諾手続等が必要な場合は受託者にて適切に対応すること。
システムの稼働地域	自動採点プログラムの運用に当たって、利用するサーバー類は全て、国内に設置されるようにすること。
ネットワークの構成	取り扱う問題情報・データの性質に鑑み、問題情報・データを実際に取り扱う者やアクセス可能な端末等の範囲を定め、その範囲でのプライベートネットワークを構築するなど、適切な対応を行うこと。
問題・解答データの取扱い	自動採点プログラムによる問題情報・データの取扱いも、あらかじめ定めた範囲内での運用とし、これらが取り扱われている最中は、外部ネットワークには接続されないようにすること。
クラウドサービスの活用	クラウドサービスの活用は、上記環境要件を満たすことを条件に、これを認める。

加えて、特に本事業で実行したような生成 AI の原理を用いる自動採点では、いわゆる AI 技術の利活用に対して、十分に注意をはらうことが求められると考えられます。

そのため、それぞれの解答データと紐づく児童生徒の個人情報に対する取り扱いや、AI の出力内容に対する分析等に対しては、例えば「経済産業省　AI 事業者ガイドライン」や「国立研究開発法人産業技術総合研究所　機械学習品質マネジメントガイドライン」のなかで取り上げられるような、現在時点で一般に留意すべき観点に則った AI の運用が求められると考えられます。

4.4.1.3　　プログラムの構築・運用・処理の要件

プログラムの構築については、セキュリティ等の要件を満たした上で、調査の実施後解答データが集められてから遅滞なく、教師データの収集や、自動採点の処理を開始できるような状態にあることが望まれます。

プログラムの運用・処理については、期間内に作業を終えることを前提とした体制が求められますが、個々の観点としては、次の表に示すようなものが考えられます。

観点	要件の整理
プログラムの処理時間	処理完了までに特に長時間を要したり、出力の信頼度を用いるなどして解答の一部のみを処理したりするような場合には、それが全体の採点工程の効率化に資することの説明が求められると考えられます。
ハードウェア・ソフトウェアの構成	１設問あたり最大 100 万行程度の解答データを処理する必要があることを念頭に、適切な規模の処理サーバー導入やプログラムの並列化など構成の工夫によって、必要時間内で処理が可能なことを示すことが求められると考えられます。
その他サービス・API 等の利用	従来の人手による採点システムで用いるサーバー等の範囲を超える、特定のネットワークサービス等が自動採点プログラムの処理中に用いられる場合には、それが不測の事態で利用できなくなった場合の予備案を示すなどして、自動採点プログラムの新規導入リスクを軽減するための方策を示すことが求められると考えられます。

4.4.2 自動採点プログラムの採点精度に関する要件

実際の採点精度は出題される問題と解答データの内容に強く依存するため、問題内容や解答類型の確定よりも前に、分類精度に関する内容を要件として定めるのは現実的ではありません。また、自動採点を用いた効率化については人手による採点プロセスとの連携とも併せて検討されるべき内容であることからも、悉皆調査として自動採点プログラムに求められる必要な分類精度を、仕様の策定時点で具体的な値として示すのは難しいと考えられます。

このため、提案時点では、実績などをもとに自動採点プログラムによって一般にどの程度の効率化が期待できるかを示したり、人手のみによる採点プロセスと自動採点を併用した採点プロセスでどのような違いが生まれるのかを示したりするなどによって、自動採点プログラムと人手による採

点を併用した場合の、全体の工程の妥当性について確認するとともに、実際に採点期間短縮の達成が見込まれるかどうかをみる手法が考えられます。

4.4.3 人手による採点との連携に関する要件

自動採点プログラムの処理は、人手による従来の採点工程の中に組み込まれると想定されることから、全体の作業工程の中で、自動採点の結果が人手による採点工程へと遅滞なく連携されるような体制となっているかどうかをみる必要があると考えられます。

また、自動採点に必要な教師データを効率よく収集する必要性が示唆されたことから、上位採点者を含めた人手による採点体制についても、引き続き重要な要件になると考えられます。

このため、従来の人手による採点体制についても引き続き評価するとともに、提案される人手による採点体制が、教師データの収集をはじめとした自動採点プログラムとの連携に際しどのようにシナジーを生むのかを、併せて評価することが求められると考えられます。

4.4.4 PBT 調査との連携に関する要件

今回の効率化の推計は特に自動採点プログラムを用いた場合であって、PBT による調査には効率化の程度が適用できないことに留意する必要があります。

このため、同等の規模であっても PBT 対象教科の採点終了時期と、CBT 対象教科の採点終了時期がずれることを考慮し、採点計画を組み立てる必要があります。具体的には、CBT で実施した教科のみを優先的に分析にあてるなどして、IRT 分析の完了時期と正答率等の単純集計の完了時期を揃えるための工夫や、教科によっては機械採点処理を行う問題と人手のみによって採点を行う問題が混在した場合に、他の教科と同時期に工程を完了させるための、採点者の管理計画を示すことが必要と考えられます。

5 まとめ

本報告では、自動採点プログラムが採点期間短縮にどの程度有効となるかを推計するために、人間による採点で多くの工数を占める記述式問題に対して自動採点プログラムを稼働させ、その結果として自動採点を適切に工程に組み込むことで、悉皆調査において CBT を活用する意義の１つである「より効率的な採点の実現」が達成可能であり、効率的な採点による採点期間の短縮によって、令和７年度理科 CBT 調査の中で新たに求められる項目反応理論（IRT）の活用までを一定期間内で完了可能であることが示されました。

併せて、解答類型や正答の条件のなかには、人間が解答の文章を読むことなく一意に定まるものも多数存在することから、言葉の意味内容を解釈するプログラムだけではなく単語の検索などによる分類手法を適切に併用することで、より精度が高く効率的な自動採点が実施可能なことも確認できました。

一方で、特に英語については教科特性に由来する特徴的な採点基準により、日本語で出題される教科にはない課題も確認されました。令和８年度全国学力・学習状況調査（悉皆調査）にて、英語に対して CBT を導入するにあたっては、英語の教科特性に対し効果的な自動採点の手法や人間による採点との適切な役割分担について、引き続き課題の検証が強く望まれます。

また、自動採点が行う内容判断の精度については、最も良い場合でも全国学力・学習状況調査の仕様として求められる品質には未だ届かないことから、精度向上のためには人間の目視による採点作業、品質確認が引き続き重要である点も見逃せません。

社会状況としても、言語モデルの開発や計算資源の増加は近年著しく、本年度の中においても日々更新されている分野です。CBT の導入を機に、結果返却の早期化へのニーズも高まっている中、採点品質を確保しつつ、自動採点を更に活用していくことが必要になってきます。100 万人規模の調査実施、採点にあたり、効率的な遂行のために人間と自動採点プログラムがどのように連携し作業を進めていくかという課題に対しては、その時点での様々な要因を考慮しながら、引き続きの調査、試行・検証が望まれます。

以上