

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8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.....	3
2	自動採点と人間を併用した採点工程の探索.....	4
2.1	令和6年度本調査研究で想定した採点工程.....	4
2.2	令和6年度本調査研究で確認された課題.....	5
2.3	採点工程の改善案の考察	6
2.4	採点工程の改善案に対する検証	7
2.4.1	検証対象とする調査問題の概要.....	7
2.4.2	記述内容以外の形式的な要素による分類手法.....	10
2.4.3	意味内容によるクラスタリングを用いた採点精度の向上.....	16
2.4.4	頻出表現や必須表現をもとにした形式的な事前分類.....	21
2.5	問題に応じた採点工程の考察.....	23
2.5.1	形式的なラベリング手法を採点工程に組み込む際の留意点	24
2.5.2	問題に応じた採点工程の考察.....	25
2.6	問題に応じた採点工程をふまえた、CBT 記述式問題に対する採点工数の推計	28
2.6.1	採点工程を最適化することに伴う、自動採点精度の変化.....	29
2.6.2	記述式問題の採点工数推計	33
3	本調査研究内テーマ B（CBT での調査実施等に関する試行・検証）の出題問題を用いた採点時間の推計.....	36
3.1	テーマ B で出題された問題の概要	36
3.2	作業時間の計測方法と留意点.....	37
3.3	選択式問題の採点時間推計.....	38
3.4	短答式問題の採点時間推計.....	39
4	令和9年度の全面 CBT 調査における採点工数の推計.....	43
4.1	工数推計の前提条件.....	43
4.2	過年度調査の構成を参考とした採点工数の推計.....	45
4.2.1	出題問題数の推計.....	45
4.2.2	出題問題内容に対する作業時間の推計.....	47
4.2.3	受検人数・有効解答件数の推計.....	49
4.2.4	総作業時間の概算と評価.....	50
5	まとめ	53

1 調査研究の目的と背景

令和9年度以降の全面 CBT 化が予定される、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、令和7年度以降の段階的な CBT 移行のなかでも「これまでの採点品質を担保できる範囲内で、従来の採点工程の一部に自動採点を活用することで、採点の効率化を実現する（本事業仕様書より抜粋）」ことが求められています。

これをふまえ、令和6年度の、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（以下、「本調査研究」と称する。）のなかで、CBT 記述式答案の採点に関する試行・検証として、自動採点プログラムを用いた場合の、記述式答案の採点精度や効率化の程度について検証を行いました。

この中で、自動採点プログラムを用いた採点効率化・採点期間短縮については一定の有用性が見られた一方で、出題される教科や問題によりその効果に大きなばらつきが見られるなど、令和8年度の中学校英語における CBT 悉皆調査や、令和9年度以降の全面 CBT 化に向けた課題が併せて確認されました。

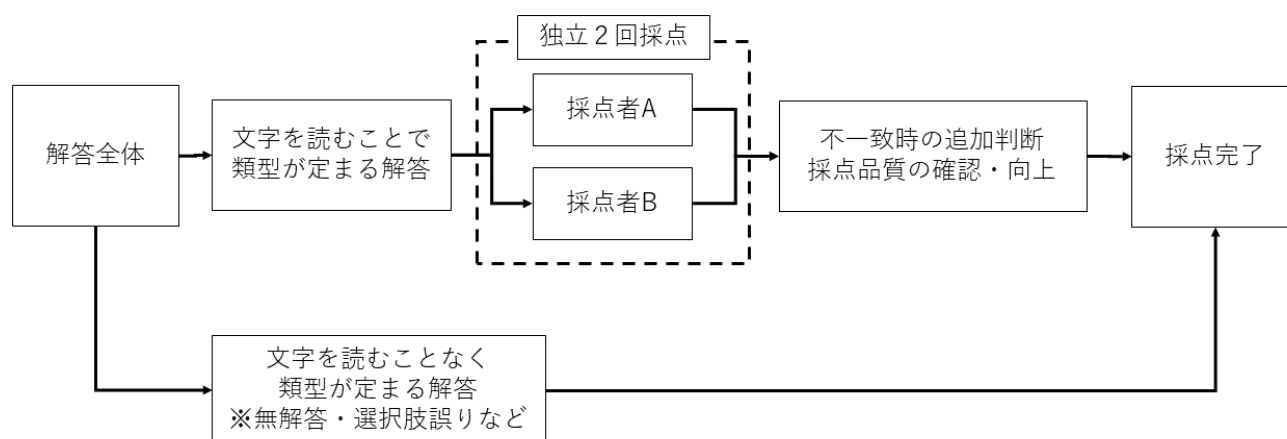
令和7年度の本調査研究では、これらの課題を解決するため、自動採点を導入した際により効果的となる採点工程の全体像について考察するとともに、令和9年度に小中学校で実施される見込みの国語、算数・数学の CBT 悉皆調査において、従来の人手のみによる採点と比較してどの程度の効率化が見込めるかを推計することを目的とした試行・検証を行います。

2 自動採点と人間を併用した採点工程の探索

本項では、令和6年度本調査研究のなかで実施した自動採点プログラムの測定・検証内容について、検証の際に前提とした採点工程と検証結果に伴う課題を整理することで、採点工程全体に対して改善が望める部分を考察し、本報告書内で測定・検証する内容を確認します。

2.1 令和6年度本調査研究で想定した採点工程

令和6年度本調査研究の中で示されている、自動採点プログラムによる測定と、作業効率化の見積では、従来人間の採点者のみで実行されていた採点工程として、下の図に示すような流れが仮定されていました。



上に示すように、無解答や選択肢部分の誤りなどを除いた、文字を読むことで類型を確定させる解答全体に対して、異なる2名による独立2回採点を行い、2名以上の判断が一致することをもって採点完了とするという採点工程です。

自動採点プログラムの実行の際にもこの工程に倣い、文字を読んで判断する必要のある解答全体を一括で扱い、全体からランダムにサンプリングされた解答群を教師データとして、生成された分類モデルを、残った解答すべてに対して同様に適用する手法を採用していました。

また、自動採点を実施した際の採点精度が各教科・問題で差が見られるかどうかを比較し、精度が変化する要因について考察するために、全教科全問題に対して、同一の自動採点プログラムや採点工程を適用したことも特徴となっています。

2.2 令和6年度本調査研究で確認された課題

令和6年度本調査研究の中で確認された、人手による採点から全面的に自動採点プログラムへと置き換えた採点工程を実施した際に考えられる課題としては、下の表に示す内容が主に示されました。

項目	課題の概要
個々の教科・問題の内容に由来する自動採点精度の振れ幅	自動採点の精度は常に一定ではなく、問題内容や解答傾向などによって振れ幅が生じる。 また、問題の中で設定される正答の条件も、教科や問題の内容によって数や基準が大きく異なり、自動採点の精度にも大きく影響する。
自由記述を認める問題で見られる、解答表現の幅広さに由来する採点難易度の上昇	特に英語の長文記述問題で見られた課題として、場面設定に由来して、他教科の記述式問題と比べて解答の範囲が広がったことで、教師データで想定しなかった記述内容に対する判断難易度が高くなったと考えられた。
1文字単位の判断を要求する採点基準	特に英語では文法や語句の正しさが評価対象となって類型が決まるため、1文字異なるだけで類型が変化する場合があり、意味内容を解釈する自動採点プログラムではそのような変化を検出する難易度が高い。

本年度の本調査研究においては、上の表に示した課題を解消あるいは軽減するための、採点の自動化、効率化の手法について考察・探索することが求められていると考えられます。

2.3 採点工程の改善案の考察

令和6年度本調査研究で確認された課題は、項番 2.1 で示したように、これまで採点者が行っていた独立 2 回採点の作業の部分をそのまま自動採点プログラムへ置き換えることを仮定し、なおかつすべての問題に対して同一の採点工程を適用したことで、各問題の内容や教科特性に依存した性能となったために生じたと推測できます。

このため、自動採点プログラムの性能に影響されない解答分類手法を考えるとともに、各問題の内容や教科特性に応じた採点工程を構築することができれば、令和6年度本調査研究で見られた課題を解消あるいは軽減することができると予想されます。

これをもとに、令和6年度の本調査研究の中で見られた課題が生じた原因と、課題に対する改善案を下の表に示します。

課題が生じた原因の考察	改善案
意味内容による分類のみでは、個々の教科や問題内容、児童生徒の解答傾向によって、採点精度に振れ幅が生じてしまう。	意味内容による分類以外にも、正誤を一定程度分類できるような形式的な判断基準を追加することによって、問題内容に由来する採点精度の振れ幅に影響されないような採点工程を構築する。
一度にすべての解答に対し分類を実施しようとすると、幅広い内容表現を含む問題に対しては分類が困難になってしまう。	意味内容による分類をすべての解答に対し同時に実施するのではなく、近い表現同士でクラスタリングを実施し、個々のクラスタに対して自動採点を実施することで、プログラムが正誤の境界を分類しやすくする。
全教科・全問題で同一の採点工程を仮定したため、自動採点の分類難易度が教科や問題内容に依存してしまった。	教科・問題が持つ特性や、解答の傾向に応じた採点工程を柔軟に構築することで、より効率的な解答の分類を実現する。

この改善案をもとに、項番 2.4 のなかで下の 2 つの分類手法について検討、測定を行います。

- ①（項番 2.4.2）：記述内容以外の形式的な条件による分類手法の探索
- ②（項番 2.4.3）：記述内容をもとにした解答群のクラスタリング手法の探索

この 2 種類の改善案を、各教科・問題の特性に応じて適切に組み合わせることで、より効率的な採点工程が構築できると推測され、教科・問題に応じた採点工程案について項番 2.5 の中で示します。

2.4 採点工程の改善案に対する検証

本項では、各教科・問題が持つ特性に応じたより効率的な採点効率化工程について検証するため、項番 2.3 の中で考察された各分類手法について測定を実施します。

2.4.1 検証対象とする調査問題の概要

本項で示す分類手法の検証対象として、令和 6 年度本調査研究で使用した問題のうち特に総工数に大きく影響を及ぼすような採点に長時間を要する記述式問題と、特に教科特性としての課題が見られた英語記述式問題を取り上げ、下の表に示す 4 問を対象とした採点工程の検討と検証を行います。

・対象問題の一覧

問題名	問題概要
国語	令和 6 年度本調査研究 項番 3.2 で、検証対象問題（解答に多様性があり、採点難易度が高い）として選定された長文記述問題
算数	
英語 A	令和 6 年度本調査研究 項番 3.5 で、経年変化調査問題に対する採点精度測定を行った問題
英語 B	※英語 A：短い文章で答える問題 ※英語 B：長文記述を求める問題

・対象問題の概略

教科	問題内容概略	正答の条件概略
国語	問いに対して設問に書かれた内容を取りあげながら、自身の考えを記述する。	①：問題文の特定箇所をとりあげて記述する ②：問いに対する自身の考えを記述する ③：指定字数で記述する

教科	問題内容概略	正答の条件概略
算数	グラフの内容を読み、推測できることを選択肢の中から選んだ上で、その根拠を記述する。 なお、正答と認められる記述手法は複数のアプローチが想定される。	①：正しい番号を選択する ②：答えを導くための着眼点 A を記述する ③：答えを導くための着眼点 B を記述する ④：根拠となる式や言葉を記述する ⑤：根拠となる答えの値を記述する
英語 A	問題の指示に対し、1 文程度の短い英文で応答する。 なお、指示事項を守る限りは、内容の自由記述は認められる。	①：問いに対して適切な内容で応答する ②：英文法の形式事項を守って記述する
英語 B	問題で与えられたテーマに対し、自身の意見や体験などを文章で発表する。 なお、解答に必要な語数の下限が定められている。	①：問いに対する自身の考えを記述する ②：英文法の形式事項を守って記述する ③：指定語数で記述する

なお、各問題の内容・正誤判断基準等の詳細については非公開であることに加え、各問題で正答の条件数や類型数、正答・準正答の区別などの設定がすべて異なることから、本項の中で行う検証では、分類基準を揃えるため、各問題・解答に対する採点結果を次の表に示す 3 種類に分け、この 3 種類をより効率的に分類するための手法について考察します。

また、無解答や選択肢の誤りなどによって、記述内容を全く読むことなく一意に解答類型が定まる答案は除外し、内容判断を要する解答のみを検証の対象とします。

・解答の分類

分類項目	分類の定義
正答	内容判断を必要とする正答の条件をすべて（または正答に充分なだけ）満たしている解答。
誤答（一部条件が不足）	内容判断を必要とする正答の条件を 1 つ以上満たすが、正誤としては誤答に分類される解答。

分類項目	分類の定義
誤答（条件の全部が不足）	内容判断を必要とする正答の条件を1つも満たさない誤答

検証対象とする4問題について、各問題の正誤分布は下の表のとおりでした。

問題名	正答	誤答（一部不足）	誤答（全部不足）
国語	25.01%	58.46%	16.53%
算数	35.16%	22.92%	41.92%
英語 A	53.46%	22.19%	24.35%
英語 B	10.21%	19.73%	70.06%

※特別な注釈のない限り、パーセンテージは小数第3位で四捨五入。以降の表も同様。

ここで、全国学力・学習状況調査における採点品質の仕様として、過年度の調査においては「サンプリングされた1,000件を確認して、正誤不一致となる解答が含まれないこと」が求められています。

つまり、通常は上の表に示す確率で正誤が分布している解答群に対して、何らかの機械的処理を行うことで、正答または誤答の可能性が非常に高い解答群を抽出することができれば、それらに対して正答または誤答のラベリングを行うことで、実質的に自動採点や人間による採点と同等の処理を行えることが期待できます。

例えば、抽出した解答群の中に正答となる解答の期待値が90%なのであれば、それらの解答群に対して全て「正答候補」とラベリングをすることで、内容の判断を実施する前から90%の精度で採点できていると扱えます。

このような機械的処理によるラベリングが、内容判断による自動採点プログラムと同等以上の精度で正誤を分類できるのであれば、これまでは採点作業が必要な解答全体を自動採点プログラムによって置き換えていたうちの一部を、形式的な機械的処理と「正誤の候補」のラベリングで置き換え、最終的な品質確認だけを人間の眼で実施することで、単に自動採点プログラムを使用する以上の効率化、省力化を期待することができます。

このため本項では「内容を読み取る以外の機械的処理によって、正誤いずれかに属する可能性が高い群を抽出する」ことを目標として、分類手法の考察と検証を行います。

2.4.2 記述内容以外の形式的な要素による分類手法

本項では、問題や記述の内容によらずに実施することが可能な、形式的な要素による分類手法について検討します。

2.4.2.1 同一解答に対する一括採点

令和6年度の本調査研究では、短答式問題において解答の種類が特定の範囲に集中しており、特に一部が頻出解答として見られることから、そのような頻出解答を上位採点者などが一括で採点することで、採点効率化が見込めることが確認されていました。

記述式問題についても同様に、文字列が一致する解答の採点結果は全く同じと考えられることから、記述式問題の中で見られた同一文字列に対しても、一括で採点する手法について検討することができます。

各問題に対して、異なる児童生徒間で文字列が完全に一致した解答の割合について集計した結果は下の表のとおりです。なお、特に英語では、大文字・小文字の区別など文字種の違いによって類型や正誤が変化する可能性があることから、ここで示す「同一解答」とは、大文字・小文字、句読点、スペース・改行などの、すべての入力完全に一致したものを同一と扱っています。

問題名	2件以上存在した同一解答が 解答全体に占める割合	2件以上存在した同一解答の 種類数
国語	1.50%	20 種類
算数	0.00%	/
英語 A	42.52%	573 種類
英語 B	3.40%	128 種類

集計の結果、特に短い英文による解答を求める問題（英語 A）では、573 種類の文字列で解答全体の約 42.5%を占める結果となりました。

解答者の能力分布が同一と仮定すると、人数が増加しても頻出解答が占める割合は大きく変化しないと考えられることから、この問題を 100 万人規模で出題した場合には、573 種類の文字列に採点結果を与えれば、約 42.5 万人分の採点が完了すると予想できます。

このように、記述式問題の形式であっても、内容としては自由記述を認める一方で、実際に解答する際には問題の指示や文中の記載内容により、解答が一定の範囲に集中することが想定できる場合があります。そのような問題であれば、頻出解答に対しては自動採点プログラムを経由させずに、特に人間の上位採点者などによって確実な採点結果を与えることで、高品質かつ少ない工数で一定範囲の採点を完了できると予想されます。

一方で、長文の自由記述を求める問題では、異なる解答者間ではほとんど解答文字列が一致しない結果となりました。文字列が一致する場合も、例えば「わかりません」のように、内容を評価できない短い記述が偶然一致していたものが大半を占めていました。

100 万人規模であれば 1%の解答が 1 万件に相当するため、そのような解答を一括で採点する効果は一定程度あると考えられるものの、基本的には、長文の自由記述問題では各児童生徒の文字入力 が完全に一致する確率は非常に低いことを前提に、採点工程を構築することが求められると推測できます。

2.4.2.2 解答字数・語数によるふるい分け

記述式問題で正答を解答するためには、正答の条件をすべて（または十分な数）満たす内容を盛り込んで記述する必要があります。このため、記述の分量が多いからといって正答とは限りませんが、記述の分量が少ない場合には、正答の条件を十分に満たす可能性が低くなり、誤答となる可能性が高くなると予想できます。

この予想をもとに、記述字数や語数によって「内容が不足している可能性が高い解答群」をふるい分けられないかを検証します。

①：字数によるふるい分け（国語・算数を例にした検証）

国語と算数の各問題について、正誤に応じた解答群の平均字数を下の表に示します。加えて「正答として通常求められる分量から著しく記述が少ないなら、内容が不足している可能性が高い」という予想の裏付けのため、達成度に応じた解答字数の標準偏差を併せて示します。

教科・正誤	国語		算数	
	平均字数	標準偏差	平均字数	標準偏差
正答	95.10 字	6.39 字	83.83 字	43.82 字
一部不足	81.11 字	20.42 字	55.90 字	25.65 字
全部不足	55.62 字	34.64 字	44.24 字	29.64 字

事前予想の通り、正答と誤答の間では平均字数に差が見られ、特に正答の条件を1つも満たさない解答では特に解答字数が少なくなる結果となりました。加えて、特に国語の各解答群が持つ標準偏差からも、「字数が少ないと正答の条件を満たさなくなる可能性が高くなる」という予想が裏付けられたことが推定できます。

なお、今回対象とした問題では、特に正答となる解答群について、国語と算数で標準偏差が大きく異なる結果となりました。これは、今回対象とした国語と算数の各問題で正答のために最低限書かなければならない内容が異なり、国語と比べて算数では正答を記述するために最低限必要な字数が少なく、正答となる字数の下限が低くなったことが標準偏差の増加につながったことが推測されます。

実際の採点工程構築では、今回測定した算数の問題のようなごく短い記述でも正答となるケースに注意する必要があるものの、基本的には、求められた内容すべてを記述しないと正答とならないような問題であれば、字数が少ない記述に対して、暫定的に「誤答の可能性が非常に高い解答群」として振り分ける採点工程を最初に加えることで、効率的に誤答を分類できると考えられます。

今回対象とした国語の問題を例に、字数を基準として形式的に振り分けた場合の、正誤分布の集まり方を下の表で示します。

項目	字数	全体に占める割合	正答	一部不足	全部不足
全体の比率	/	100%	25.01%	58.46%	16.53%
正答の平均以上	96 字以上	33.65%	46.50%	46.01%	7.49%
一部不足の平均以下	81 字以下	31.39%	2.91%	63.07%	34.04%
全部不足の平均以下	55 字以下	15.19%	0.16%	47.87%	51.97%

上の表では、今回対象とした国語の問題について、例えば 55 字以下の解答に対し「誤答の可能性が非常に高い解答群」というフラグを与えることで、全体の 15.19%に対し、正誤一致率 99%以上で事前分類できることが示されており、81 字以下（全体の 31.39%）に拡大しても正誤一致率 97%以上で事前分類できることになります。

逆に、十分な記述量がある解答群では、正答の可能性は全体平均よりも高くなるものの、正誤のどちらも同程度に含まれており、形式的な分類の効果としては小さいことが推測されます。

この結果からも、「記述量が多いからと言って正答になるとは限らないが、記述量が少ないと誤答の可能性が非常に高くなる」という予想が裏付けられ、自動採点と字数による分類を併用する手法として、例えば字数がある程度多い解答のみを対象として教師データの作成や自動採点プログラムの実行を行い、記述量が著しく不足している解答群に対しては、「誤答の可能性が高い」とラベリングした上で、人手による品質向上作業などによって実際の類型を確認する、といった工程が考えられます。

②：語数によるふり分け（英語を例にした検証）

英語については、一般に字数よりも単語数によって文章の構成要素が評価されることから、解答の語数と正誤の関係性について検討・測定します。

加えて、特に英語の採点基準としては、文法についても正しく記述することが一般に要求されることから、解答が文として成立する最低語数に満たない場合、正答となる可能性は著しく低くなることが予想できます。

この予想をもとに、英語の各問題における、達成度に応じた平均解答語数と標準偏差について下の表で示します。

なお、長文記述を求める問題（英語 B）については、項番 2.4.1 の表でも触れたとおり要求単語数に下限があり、それに達していない解答に対しては内容判断を行わず所定の類型に分類されました。このため、下の表で示す平均語数については、語数の条件を満たし、類型分類のために記述の内容判断が行われた解答のみを対象として集計しています。

教科・正誤	英語 A		英語 B	
	平均語数	標準偏差	平均語数	標準偏差
正答	7.59 語	3.78 語	28.93 語	5.81 語
一部不足	5.71 語	2.38 語	27.93 語	5.48 語
全部不足	2.57 語	2.02 語	27.35 語	3.95 語

短い英文で解答を求める問題（英語 A）では、正答と誤答の間で平均語数の差が大きく見られ、特に 1 文として成立する語数にも満たない場合、ほとんどは正答の条件を全く満たさないと推測することができます。

一方で、長文で記述する問題（英語 B）では、内容部分の正誤による解答語数の差はほとんど見られませんでした。これは、前述の通り設問に語数指定が求められており、内容判断の対象となった解答を記述した生徒については「問題文中で示された語数指定を守ろうと解答する」という要素が影響して、短い文章による誤答が少なくなるという要因が、状況としては考えられます。

過年度調査の公表問題を例にとると、令和5年度・平成31年度（令和元年度）の全国学調（それぞれ中学校英語 大問10）では、一定の語数に満たない英文については、特定の類型に分類される基準となっていました。

しかしながらこのような場合も、語数が事前の類型分類の手法として機能していると考えられ、英語においても、記述量が少なければ誤答となる可能性が高くなるという予想は成り立つことが推測できます。

今回の検証で特に正誤間で語数に差が見られた、短い英文で解答を求める問題（英語A）について、語数を基準として形式的に振り分けた場合の正誤分布の集まり方を下の表に示します。

項目	語数	全体に占める割合	正答	一部不足	全部不足
全体の比率	／	100%	53.46%	22.19%	24.35%
正答の平均以上	8語以上	13.07%	83.51%	13.29%	3.20%
一部不足の平均以下	5語以下	38.45%	9.66%	34.92%	55.42%
全部不足の平均以下	2語以下	9.49%	0.00%	0.68%	99.32%

無解答を除いた解答のうち約9.5%は2語以下であり、2語以下の解答の99%以上は正答の条件を1つも満たさないという結果となりました。本問題であれば、2語以下の入力で解答されたものに対しては、1文にも満たない記述とみなして一律に誤答とラベリングするだけでも、自動採点よりも高い精度で分類できたことになります。

また、英語以外の教科においても同様に、文章として成立する最低限にも満たないようなごく短いテキストに対しては記述不足のラベルを与え、品質向上作業によって実際の類型を確認する採点工程が考えられます。

2.4.3 意味内容によるクラスタリングを用いた採点精度の向上

令和6年度の本調査研究で見られた課題として、特に英語の長文記述問題で、解答内容の幅広さから、自動採点プログラムが内容を解釈する難易度が高くなったのではないかと指摘が挙げられていました。

この課題に対し本項では、解答全体から特定の表現のみを抽出して、抽出した解答群の中で教師データ作成や自動採点処理を実施する、クラスタリングの手法による採点精度の向上について検証します。

2.4.3.1 クラスタリングの条件を推定できる場合の自動採点

出題される問題内容によっては、児童生徒が記述する表現を事前に予測できる場合があります。

今回検証の対象とした英語の長文記述問題（英語 B）では、問題文の中で解答テーマの例が単語として例示されており、出題内容としては自由記述を認めている一方で、記述される解答としては、問題中に示されている単語を用いる可能性を強く推測することが可能でした。

実際に集められた解答としても、解答全体で使用されている単語を集計した結果、問題中で例示されていた3つの単語を含む解答が大半を占めており、本項ではこの頻出単語をクラスタリング条件として、特定のテーマ内で教師データの作成や自動採点を実施した場合の精度向上について測定・検証を行います。

クラスタリング条件となる3つの単語の含まれ方と、各クラスが全体に占める割合は下の表のとおりです。

クラスタリング条件	全体に占める割合
単語 A を含む	47.71%
単語 A を含まず、単語 B を含む	23.09%
単語 A,B を含まず、単語 C を含む	21.81%
単語 A～C のいずれも含まない	7.39%

※複数の単語を同時に含む場合は、より大きいクラスに分類した。

90%以上の解答では問題の中で示された単語をそのまま使用しており、ほとんどの児童生徒は、問題文の中で示された内容に沿って解答を作成していることが推測できます。

このうえで、頻出単語 A～C を条件として分割した 3 つのクラスタそれぞれに対して自動採点を実施し、採点精度が向上するかどうかを確認しました。

測定条件は令和 6 年度の本調査研究と同様とし、令和 6 年度本調査研究で使用したものと同等の、機械学習による分類を実施する自動採点プログラムを使用し、各クラスタ別に教師データを 1,000 件ランダムに抽出して自動採点を実施しました。

これを令和 6 年度本調査研究と同様に 5 回繰り返し、自動採点結果について、採点済みの結果と照らした類型一致率の平均を取ったものを自動採点の精度として、下の表に示します。

クラスタの内容	全体に占める割合	類型一致率の平均
全体を同時に自動採点した場合 ※令和 6 年度本調査研究より	100%	55.17%
頻出単語を含むクラスタの平均	92.61%	61.81%
単語 A を含むクラスタ	47.71%	60.49%
単語 B を含むクラスタ	23.09%	62.12%
単語 C を含むクラスタ	21.81%	64.37%
その他のクラスタ	7.39%	/

頻出単語をもとにしたクラスタリングを実施して、クラスタ別に自動採点を実施することで、自動採点による類型一致率の平均が 6 ポイント以上向上しており、クラスタリングの効果が一定程度確認できました。

意味内容を解釈する自動採点プログラムでは、文法や誤字に由来する類型の分類難易度は高く、これによる類型一致率の低下が要因としては残りますが、これはクラスタリングを実施していないときも同様であることから、クラスタリングによって、解答の意味内容を解釈する難易度としては低下しており、これが精度向上につながったと考えられます。

ただし、本手法を用いるためには、分割したクラスタの数だけ教師データの作成が必要である点に留意する必要があります。

本項の検証では、3つのクラスタに対して1,000件の教師データを作成しているため、本来1,000件であった教師データを、クラスタリングを併用して自動採点を実施する際には3,000件作成しなければならないことを示しています。

実際の全国学調では、教師データは人間が新たに作成すると想定され、今回を例にとると教師データ作成の工数が想定の3倍必要となることから、教師データ作成の日数や上位採点者等の作業者が充分かどうか検討したうえで、採点工程にクラスタリングを組み込むことが必要と考えられます。

2.4.3.2 記述内容をもとにした自動クラスタリングによる分類

国語のように同義語や送り仮名のバリエーションが多い問題や、特定の語句を書く指示がない問題では、項番2.4.2で実施したような、文字列に対する一致検索でクラスタリングの条件を設定することが困難な場合があります。

このような場合には、各解答をベクトル埋め込みなどで数値変換した上で、k-means法を始めとした自動のクラスタリング手法を適用することによって、自動的に近い意味の内容を同じクラスタにまとめる手法が考えられます。

今回検証で用いた英語の長文記述問題を例にとると、自動クラスタリングの実行によって、頻出単語が事前に不明な場合でも、同じ頻出単語が同じクラスタに集まることで、項番2.4.3と同様に、クラスタ別の自動採点が可能と考えられ、本項では実際に頻出単語を同じクラスタに自動で集めることができるかどうかを検証します。

また、事前に頻出単語が不明な場合には、いくつかのクラスタに分かれるかも推測困難となることから、様々なクラスタ数を指定してもクラスタリングが正しく機能するかを併せて確認します。

項番2.4.3で対象とした英語記述式問題について、自動クラスタリングを実施した場合の、各クラスタに含まれる解答内容の分布は下の表のとおりです。

・ 4 クラスタの場合 ※頻出単語の予想から「単語 3 種類+その他の内容」と想定した場合

クラスタ番号	最も多く 含まれた単語	左列の単語が 含まれる割合	解答全体に対する クラスタの大きさ
1	単語 A	96.92%	37.98%
2	単語 B	98.05%	22.88%
3	単語 C	95.05%	22.80%
4	その他	／	16.33%

・ 5 クラスタの場合 ※頻出単語を多く見積もり「単語 4 種類+その他の内容」と想定した場合

クラスタ番号	最も多く 含まれた単語	左列の単語が 含まれる割合	解答全体に対する クラスタの大きさ
1	単語 A	96.34%	22.99%
2		95.66%	19.65%
3	単語 B	98.22%	22.80%
4	単語 C	95.51%	22.50%
5	その他	／	12.05%

・ 8 クラスタの場合 ※「単語 3 種類+その他の内容」を 2 倍に細分化しようとした場合

クラスタ番号	最も多く 含まれた単語	左列の単語が 含まれる割合	解答全体に対する クラスタの大きさ
1	単語 A	95.07%	15.70%
2		98.45%	15.28%
3		93.15%	13.22%
4	単語 B	97.40%	12.13%
5		97.96%	10.92%
6	単語 C	95.40%	12.85%
7		93.02%	10.16%
8	その他	／	9.73%

今回の測定結果としては、頻出単語の予想ができない場合でも自動クラスタリングを実行することによって、概ね 95%以上の精度で、特定の単語を含む文章を同じクラスタに分割することができました。また、より多くのクラスタ数を指定したような場合であっても、頻出単語を含むクラスタの中で細分化が行われることで、全体の精度を大きく損なうことなく、同じクラスタには概ね同じ表現が含まれることを確認しました。

事前に頻出表現が不明な場合であっても、例えば設定された解答類型と等しい数を指定するなど、十分に大きなクラスタ数を設定して自動クラスタリングを実施することで、クラスタ別の自動採点が可能となって、自動採点の分類精度をより高めることが期待できます。

また、人手による採点の場合でも、特定の表現が多く含まれた 1 つのクラスタに限定した採点作業を実施することで、似たような意味内容の判断だけに集中させて、能力の習熟を図るような手法も考えられます。

ただし、項番 2.4.3 でも触れたように、クラスタリングを用いた自動採点精度の向上では、各クラスタに対してそれぞれ異なる教師データを作成することによる工数増加に留意する必要があります。特に事前の解答傾向が予想できず自動クラスタリングに頼る場合、過度にクラスタを細かく分割してしまうと、教師データ作成等で非常に多くの作業工数が割かれる点にも注意が必要です。

また、令和 6 年度の本調査研究でも測定されたように、問題文の内容を含むよう指示がある場合など、正答の条件を満たすために記述すべき内容がある程度定まっている場合には、クラスタリングを実施せずとも一定程度の自動採点精度が達成できる場合があります。

これ以外にも項番 2.4.2 で述べた手法などによって、内容の判断に頼らずに高い分類精度が達成できる場合も考えられることから、特に教師データ作成の工数増につながるクラスタリングを実施するかどうかについては、問題や解答の傾向に応じ柔軟に設定することが重要だと考えられます。

2.4.4 頻出表現や必須表現をもとにした形式的な事前分類

項番 2.4.3 で用いたように、解答全体で頻出表現が予想できる場合、クラスタリングすることによって内容判断の精度を高める採点工程を考えることができますが、さらに問題の特性によっては、特定の表現の有無が正誤と強く結びついている場合があります。

例えば問題の中で必須語句が指定されている場合、必須語句が含まれない解答は内容を判断せずとも誤答となる可能性が高くなると考えられ、このような場合には、内容判断の精度向上のためだけではなく、項番 2.4.2 で取り上げたような形式的な分類手法として、頻出表現や必須表現を検索対象とした事前分類を、採点工程の中に組み込むことが検討できます。

このため本項では、特定の表現を検索条件とした正誤の事前分類手法について確認・検証します。

①：頻出表現を予想できる場合のふり分け（英語を例にした検証）

問題によっては、指示としては自由記述を認める一方で、実質的には教科書に掲載されている表現を用いるなど、特定の表現を記述することを念頭に置いて出題される場合があります。

例えば英語の場合であれば、実際の英文としては許容される単語や表現は多数考えられるものの、一般的には教科書などに記載されている表現や語句を用いて解答することが予想され、正答を記述する場合も同様に、その多くは想定正答とほとんど同一の表現を用いて解答されると予想できます。

この予想をもとに、英語 A（短い文章で答える問題）を対象として、文章の主な構成要素となる主語と動詞までが想定正答と一致しているかどうかで分類した場合の、正誤別の分類結果を下の表に示します。

解答の特徴	全体に占める割合	正答	一部不足	全部不足
全体の比率	100%	53.46%	22.19%	24.35%
主語と動詞が 想定正答と一致している	51.37%	89.67%	9.62%	0.71%
主語または動詞が 想定正答と一致しない	48.63%	15.20%	35.47%	49.33%

事前予想の通り、全体の半数程度は想定正答と同じ主語と動詞を用いて解答されており、加えて、主語と動詞が想定正答と一致するかどうかで、正答を期待できるかどうかが大きく異なる結果となりました。

項番 2.4.2.1 では文字列が完全一致している解答のみについて着目しましたが、本項のように解答の一部分が一致している解答を分割することによっても、例えば人手による作業において、文章の後半部分の違いにだけ着目して単語や表現が正しいかどうかを判断させる作業が考えられ、これによって採点の際に着目すべき箇所を減らし、作業負荷の軽減や判断ミスの減少につなげることが期待できます。

また、主語や動詞が想定正答と一致しない解答群の中でも「想定正答とは異なる頻出表現」を同様に考えることができることから、これらの頻出表現についても形態素解析や文章の前方一致検索などの手法によって抽出して、それぞれの表現別に採点を行う手法が考えられます。

②：正答のための必須表現を事前に予想できる場合のふり分け（算数を例にした検証）

問題によっては、①で検索条件とした「一般的な正答を書く際に用いられる表現」より強い条件として「書かれていなければ誤答となることが予想される表現」を考えることができます。

算数を例にとると、解答の中で計算を要求するような問題であれば、通常は解答の中に数字や数学記号が含まれると考えられ、裏返すと数字や数学記号が含まれていないのであれば、正答となる可能性が低くなると予想することができます。

この予想をもとに、対象とした算数の問題について、解答に数字を含むか含まないかによって分類した場合の、正誤別の分類結果を下の表に示します。

解答の特徴	全体に占める割合	正答	一部不足	全部不足
全体の比率	100%	35.16%	22.92%	41.92%
解答に数字が含まれている	83.67%	42.03%	23.26%	34.71%
解答に数字が含まれていない	16.33%	0.00%	21.15%	78.85%

内容を判断する必要のある解答のうち、全体の約 16%で解答に数字が含まれておらず、数字が含まれていない場合はすべてが誤答という結果になりました。

このような場合には、例えば解答中に計算を求めるような問題であれば、数字が含まれている解答のみを自動採点など通常の採点工程に割り当て、数字が含まれていない解答は暫定的に「誤答の可能性が高い解答」としてラベリングして、後の工程で実際の類型を確認するような工程を考えることができます。

また、算数・数学以外の教科であっても、問題文の中に登場する表現や問われている知識・技能に紐づく単語のように、正答を記述するための必須語句や、それに準じた単語・表現を予想することができる場合があります。

そのような場合も同様に、「正答に必要な表現が含まれた解答」と、「正答に必要な表現が含まれていない解答」にふるい分けてそれぞれ異なる採点工程を適用することで、目視確認の優先度が高い解答に人手を割くといったような採点工程が考えられます。

2.5 問題に応じた採点工程の考察

項番 2.4 の中で取り上げた工程の改善案は、すべての教科・問題に対して適用可能かどうかを逐一考えることが可能です。しかしながら、解答を細かく分割し、分類のために多くの工程を挟むほど、人間が行わなければならない作業が増加することにもつながります。

このため、項番 2.4 で取り上げたような分類手法すべてを全問題で一律に適用するのではなく、問題内容や予想される解答傾向に応じ適切に取捨選択して採点工程に組み込むことで、より効率を高めることが期待できると推定されます。

本項では、項番 2.4 の内容をもとに、形式的な処理と内容判断を組み合わせた効果的な採点工程について考察を行います。

2.5.1 形式的なラベリング手法を採点工程に組み込む際の留意点

項番 2.4 のなかで高い分類精度が得られた観点は、特に誤答に対するもので顕著でした。

これは一般に「正答に必要な記述があるからといって正答とは限らないが、正答に必要な記述がなければ通常は正答にならない」という評価によって一定程度予測が可能です。

特に項番 2.4.2 で確認したように、単語数・文字数や、算数・数学における数字や数学記号の有無といった、通常その問題に対する解答として求められる記述が不足している場合には、人間の作業者や自動採点プログラムによる独立 2 回採点から置き換え可能な精度で、事前に「誤答の可能性が高い」と想定する事が可能です。

ただし、これはあくまでも形式的な処理によるラベリングであって、実際に正誤や類型を決定するにあたっては、最終的には文章の内容を実際に読まなければ断定できない点に注意が必要です。

語数や必須語句が明らかに不足しているような場合などでは、目視確認の工程について省力化も期待できますが、それであっても、例外的な正誤不一致解答が本当に含まれていないかどうか、工程の最終段階では人間によって行う、全体の品質検収をもって仕様で求められる採点品質を達成していることを確認する必要がある点には、特に留意しなければならないと考えられます。

また、出題される問題によっては、今回確認した以外の観点を組み合わせて、正答や誤答に対する形式的なラベリングが可能となる可能性は生じますが、その場合にも、「人間やプログラムによる独立 2 回採点から置き換え可能と充分期待できる」精度で行うことが求められると考えられます。

このため、次項 2.5.2 で紹介する工程例や各種図表については、それら形式的な処理による採点精度が、人間やプログラム等と比較しても、工程の置き換え後にも充分仕様に求める品質が担保できると期待できる設定によってなされていることを前提として示します。

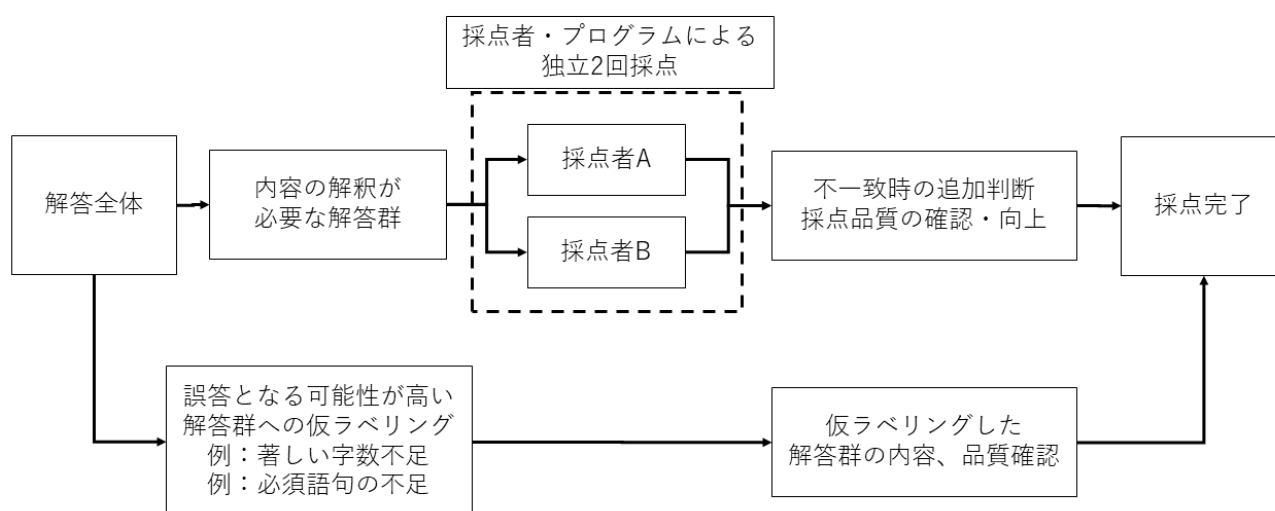
2.5.2 問題に応じた採点工程の考察

以下に具体例として、典型的な問題内容・解答傾向に対して考えられる採点工程を、次の①から③の3例で示します。

- ①：児童生徒の記述内容が一定の範囲に集まることが想定できて、自動採点プログラムによる内容解釈の精度が一定程度期待できる場合

自動採点プログラムによる内容解釈の精度が一定程度期待できる場合には、令和6年度の本調査研究で仮定したような採点工程でも十分な効果が予想されますが、それに加えて、字数が極端に少ない解答や必須語句が不足している解答のような「内容不足で誤答となる可能性が高い解答」をふるい分けた上で、「内容判断の必要性が高い解答」に集中して教師データの作成や自動採点を実施して、全体としての分類精度をさらに高めるような方法が考えられます。

これを図示すると、下に示すような採点工程が想定されます。



※無解答などの「文字を読むことなく類型が一意に定まる解答」については図から省略している。

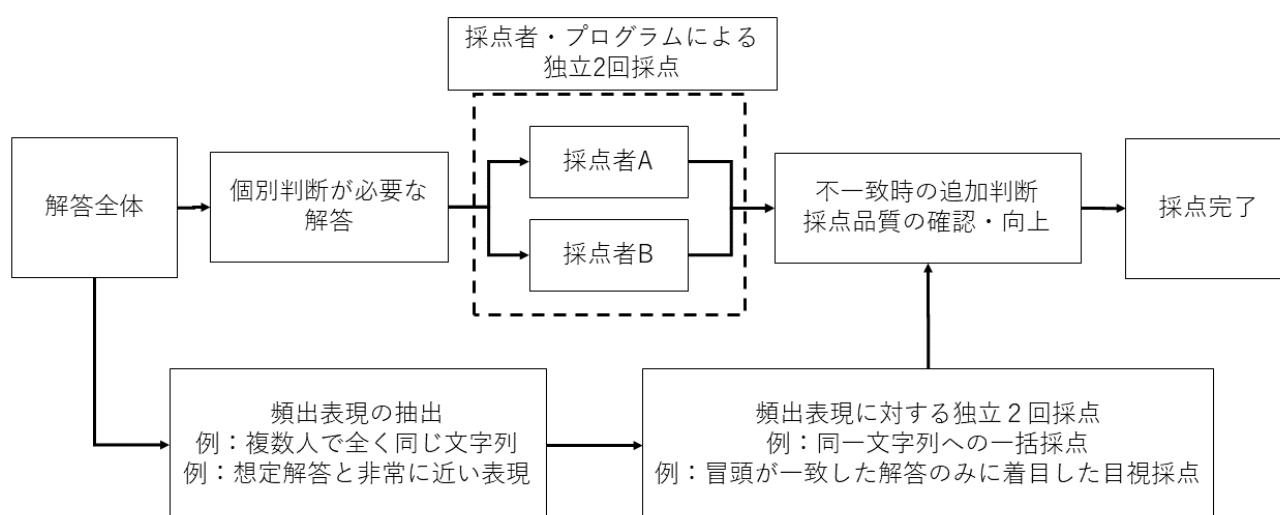
本項で例示する図は以降すべて同様

②：短い文章で解答する問題のように、記述式であっても典型的な解答、表現が想定できて、特定の頻出表現に解答が集中すると予想できる場合

記述式であっても頻出表現が解答の多くを占めると予想できる場合には、令和6年度本調査研究の短答式問題に対して確認された内容と同様に、重複・類似した解答を高速で判断するための手法が考えられます。

完全な同一文字列を一括で採点する他にも、例えば英語であれば主語や動詞の部分が一致している解答のみを抽出して採点者に判断させたり、逆に個別判断が必要な解答群に対して、採点者による目視採点の工数を集中させたりする工程を考えることができます。

これを図示すると、下に示すような採点工程が想定されます。

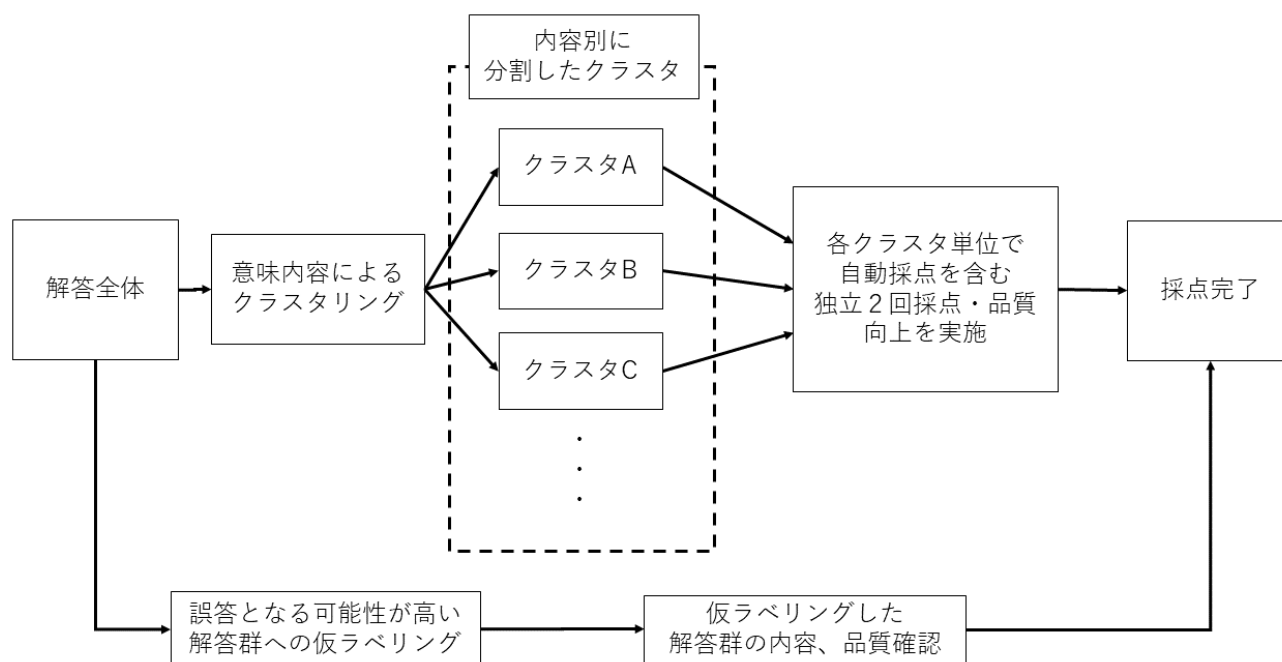


③：自由なテーマで解答することを認める問題のように、典型的な解答や表現を想定することが困難で、内容判断の難易度が高くなる懸念がある場合

特に多様な解答が想定され、そのままでは採点する難易度が高いと見込まれる場合には、クラスタリングによって似た意味内容を持つ解答群に分割してから、似た意味内容の中で自動採点や目視採点を実施することで、判断の負担を軽減する工程が考えられます。

これに加え、他の記述式問題で行うような、記述不足の解答などに対するふり分けを併せて実施して、内容の判断が必要な解答の総数を減らすことも必要と考えられます。

これを図示すると、下に示すような採点工程が想定されます。



以上3つの例ではいずれも、問題や解答の特性上形式的に分類可能な要因を切り分け、自動採点プログラムや採点者による目視採点を、解答の意味内容に対する判断や採点結果の確認といった、より品質を向上させるための作業に集中させることで効率化を図るという共通点があります。

本報告で検証した以外の問題や、今後全国学調の中で出題される問題についても同様に、形式的に処理可能な要素と意味内容の解釈が必要な要素を適切に分割した採点工程を構築することによって、自動採点プログラムや採点者がもつはたらきをより引き出すことが期待できます。

2.6 問題に応じた採点工程をふまえた、CBT 記述式問題に対する採点工数の推計

項番 2.5 で考察した採点工程案をもとに、令和 6 年度試行検証で使用した小学校国語の記述問題を対象に、令和 6 年度内での推計からどの程度の改善が見込めるかを考え、これをもとに令和 9 年度の本体調査で出題される記述式問題の採点に必要な作業時間を推計します。

令和 6 年度の試行検証では、使用した小学校国語の記述式問題（最も採点が困難と見込まれた問題）について、上位採点者に相当する採点者が目視で採点作業を行った場合の実測値として、下の表の通り採点事業者より提供を受けています。

観点	実測値
1 件あたりの採点秒数	29.27 秒
通常の採点で判断を保留した解答の割合	16.99%
判断した解答に対して誤分類が生じる割合 ※品質確認時の判断を含む	4.46%

この値をもとにした令和 6 年度の試行検証事業では、1 問あたりに必要な採点工程と各作業時間について、下の表に示すようにまとめられました。

観点	概要
採点作業手順	自動採点により 1 回採点後、人間による 2 回目の採点を行い、その後品質を確認する。 判断保留・不一致等が生じた場合の採点結果上書きは、上位採点者の合議によって全て正しく反映されると仮定する。
作業に応じた採点秒数に対する補正	・品質確認のために必要な作業時間は、1 件あたり通常の半分と見積もる。 ・保留・不一致解答に対する結果の上書きについては、1 件あたり通常の 2 倍の時間と見積もる。
1 件あたりの作業秒数計	70.02 秒
作業完了後の誤分類混入割合	0.052% ※要求品質を鑑み、完了後の誤分類率のみ小数第 3 位まで表記

本項では、これらの推計根拠を同様に用いて、項番 2.5 に示すような採点工程案によって記述式問題を採点した場合に必要な作業時間を推計し、昨年度の推計結果と比較します。

なお、本項で示すような採点工程を新たに採ることで、判断に要する時間や採点品質が各作業工程で変化する可能性はありますが、本項での推定では令和 6 年度試行検証で使用した値を再度使用して、工程の最適化による改善効果を測るものとしします。

2.6.1 採点工程を最適化することに伴う、自動採点精度の変化

項番 2.5 の④で示したように、記述式問題においては、「誤答となる可能性が高い解答群」を事前に取り分けておくことで、より効率的な採点が期待できることを示しました。

これにより、自動採点を適用する範囲についても、令和 6 年度試行検証事業における考え方では解答全体に対して教師データの抽出と分類を実施していたところを、個別判断が必要な解答に範囲を絞って自動採点を実施することで、より重要な部分に対して自動採点の効果を適用できるようになると予想できます。

本項では、令和 6 年度に実施していた自動採点の結果と、形式的な条件をもとに解答群を取り分けたあとの解答に対する自動採点結果を比較して、採点工程を問題ごとに最適化した際の、自動採点への影響について測定します。

なお、本項で使用した自動採点プログラムおよび測定方法は、令和 6 年度試行検証事業の中で精度測定のために用いられた、LLM（大規模言語モデル）による、正答の条件別に二値分類を実施するプログラムを同様に使用し、対象となる解答群からランダムに抽出した解答を教師データとして、分類モデルの生成などを実施しています。

2.6.1.1 字数が十分な解答のみを抽出した自動採点の精度

項番 2.4.2 で示したように、一般的に記述式問題の場合は正答の条件を満たすために十分な記述量が求められることから、著しく字数が不足している解答は、誤答となる可能性が相対的に高くなると考えることができました。

このため本項では、特に字数が十分な解答のみに着目した場合の自動採点について測定し、その結果を下の表に示します。

項目	令和6年度測定結果 ※解答全体が対象	本年度測定結果 ※字数が十分な解答のみ対象
平均値	79.31%	79.30%
最大値	81.28%	79.91%
最小値	77.33%	78.45%

字数が十分な解答のみに着目して自動採点を実施した測定の結果、類型一致率の最小値と最大値の振れ幅は小さくなった一方で、全体の平均値としてはほとんど変わらない結果が得られました。このように範囲が変化した理由としては、文字数の大小で事前に選り分けたことで、自動採点の対象となった解答内容の傾向が変化したことが推測されます。

字数不足のものを正答の条件不足の可能性が高いとして選り分けたことで、字数が充分書かれている解答群は、各正答の条件についてある程度の記述が含まれている可能性が高くなっています。同時に、例えば「わかりません」といったような、正答の条件と無関係な内容が書かれた解答を含む可能性が低くなっているとも予想できます。

この解答群から教師データを抽出して分類を実施しているため、どのように教師データを抽出しても一定程度採点したい解答と近しい内容記述から学習モデルを構築できていると予想され、教師データ内容の違いによって生まれる精度の振れ幅が小さくなったと考えられます。

一方で、自動採点対象としている解答群には正答と無関係な内容が少なく、相対的により小さな表現の差に対する判断が求められることになるため、採点判断の難易度としては上昇しており、見た目上の精度数値としては、選り分けずに自動採点した場合と同程度になったと推測されます。

しかしながら、昨年度字数で選り分けずに自動採点した際には対象として含まれていた、ひと目で誤答と判断できるような無関係な解答を含まずに同程度の自動採点精度が得られた、とも評価できることから、項番 2.4～項番 2.5 で示したようにあらかじめ解答を選り分けてから自動採点を実施する手法には、自動採点プログラムの機能という面でも、一定程度の効果が期待できると推測できます。

2.6.1.2 自動採点プログラムの「自信度」を用いた分類

プログラムによって「自信度」「信頼度」「評価値」など様々な呼称がありますが、一般に多くの機械学習モデルでは、推論結果（本試行検証における採点結果）とともに、その推論結果を出力する根拠となった数値を取得することができる場合があります。（※本文では「自信度」と呼称）

数値の出力工程や、個々の値に対する解釈の手法は各プログラムの内容によって異なりますが、一般には数値が高い（目的の方向に近い）ほど、より信用できる推論結果である、という点で共通しており、本試行検証で用いた自動採点プログラムについても、採点結果と同時に得られた自信度の大小によって、自動採点の精度がどう変化するかを確認します。

出力された自信度によって、自動採点結果を上位 50%と下位 50%の 2 群に分割した場合の、それぞれでの類型一致率は下の表の通りです。

自信度による しきい値	設定しない	上位 50%	下位 50%
平均値	79.30%	91.82%	66.77%
最大値	79.91%	92.90%	68.59%
最小値	78.45%	90.29%	64.40%

表で示されるように、出力される自信度が高い群では 90%以上の類型一致率が得られた一方で、自信度が低い群の類型一致率は最大でも 70%以下と、極端に分かれる結果となりました。

今回の問題を人間が 1 回採点したときに、約 17%で判断保留、約 4.5%で誤分類が生じるという実測値と照らすと、特に自動採点結果の自信度の高いものでは、十分な採点品質が得られると想定することができます。

このような結果になった原因として、教師データをもとに推論しやすく自信度が高く出力されるような解答は、教師データと近い意味内容を持っている可能性が高いことが推測されます。

例えば、教師データと 1 単語だけ異なるような解答に対する分類であれば、自動採点プログラムが生成した分類モデルに一致しやすくその自信度も自然と高くなる一方で、教師データに全く含まれない単語や文章に対する分類の場合には、分類モデルとの適合も弱く、自信度や採点精度が低下しやすくなると予想できます。

このため、自動採点プログラムから自信度が出力可能であって、それが採点精度と一定の相関を持つと期待できる場合には、積極的に利用していくことが必要になるとともに、教師データ収集の際にも、類型反応率の割合に応じてデータを与えるなどして、実際の解答群に近い意味内容を学習させたりするような、より良い学習モデルの生成を狙うための工夫も考えられます。

また、このような目で見えてわかる結果が得られた背景としては、今回学習の対象とした問題と採点結果の組み合わせが、意味内容によって判断される性質を持っていて、特定の単語の有無や誤字脱字などの形式的な表記によって判断されにくかったためと推測されます。

昨年度試行検証の中学校英語でも見られたように、意味内容としてはほとんど同じであっても、1文字・1単語の差によって正誤や類型が変化する場合には、意味内容からの採点判断では「自信を持って分けられる」ケースが少なくなってしまうと考えられ、この点でも、項番 2.4 までは述べたような形式的な観点による分類と、意味内容をもとにした評価による分類をうまく組み合わせる必要性がみられます。

なお、今回の集計ではしきい値を上下 50%で選り分けましたが、このしきい値についても、問題の特性によって最適化することで、より効率的な採点が可能と予想できます。

ただし、しきい値を厳しくすると自動採点される解答が少なくなり、逆にしきい値を広げると採点精度が平均化されてしまうという背反関係にあることから、各問題によって最適なしきい値は、実際の自動採点の実行結果や調査全体としての採点計画などから、個別に判断する必要があると考えられます。

2.6.2 記述式問題の採点工数推計

項番 2.4.2 で示した字数をもとにした分類と、項番 2.6.1 で示した事前の自動採点の結果によって、記述式の解答群を大きく 3 つに分類することができます。

解答群	作業内容	占有率	類型一致率
字数が極端に不足していて、誤答となる可能性が非常に高い	品質確認のみを行う	15.19%	51.97%
字数が一定程度ある解答のうち、自動採点結果の自信度が高い	2 回目の独立採点と、品質確認を行う	42.41%	90.29%
字数が一定程度ある解答のうち、自動採点結果の自信度が低い		42.41%	64.40%

この分類後の解答群それぞれに対し、人間が行う採点作業工程で発生する作業時間を見積もり、それらを合算することで、解答全体を採点するのに必要な作業時間を推計します。

①：字数が極端に不足している解答の品質確認

工程	作業対象の延べ割合	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
事前分類	15.19%	7.89%	7.30%	--
人手による品質確認①	15.19%	7.54%	0.33%	7.32%
人手による品質確認②	7.87%	7.21%	0.015%	0.65%
誤分類疑いで検出された解答に対する最終判断	7.97%	7.97%	--	--
作業後の採点結果	--	15.18%	0.015%	--

②：自動採点の自信度が高い

工程	作業対象の 延べ割合	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
自動採点結果	42.41%	38.29%	4.12%	--
人手による２回目採点	42.41%	30.36%	0.15%	11.89%
採点後の品質確認	30.52%	29.01%	0.007%	1.50%
判断保留・誤分類疑いの 解答に対する最終判断	13.39%	13.39%	--	--
作業後の採点結果	--	42.40%	0.007%	--

③：自動採点の自信度が低い

工程	作業対象の 延べ割合	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
自動採点結果	42.41%	27.31%	15.10%	--
人手による２回目採点	42.41%	21.66%	0.56%	20.19%
採点後の品質確認	22.22%	20.69%	0.025%	1.50%
判断保留・誤分類疑いの 解答に対する最終判断	21.69%	21.69%	--	--
作業後の採点結果	--	42.38%	0.025%	--

①～③の合計

作業工程	作業対象の 延べ割合	1 件あたり 作業時間	平均した期待作業時間 ※延べ割合×作業時間
通常の採点作業	84.82%	29.27 秒	24.83 秒
品質確認	75.79%	14.64 秒	11.10 秒
上位採点者の最終判断	43.05%	58.54 秒	25.20 秒
1 件あたり期待作業時間の合計		61.13 秒 ※令和 6 年度：70.02 秒	
100 万件あたり作業時間の推計		16,981 時間 ※令和 6 年度：19,450 時間	
作業完了後の誤分類率		0.047% ※令和 6 年度：0.052%	

これらの表より、問題内容や解答の特性に応じて採点工程を最適化することで、単に自動採点プログラムを実行する場合と比べて、作業時間を 10%以上短縮したうえで、同程度の採点品質で完了できることが期待される推計結果が得られました。

また、100 万件あたりの作業時間として 16,981 時間という値が得られ、項番 4 で述べる令和 9 年度の本体調査で出題される記述式問題に対する作業時間の推計にあたってはこの結果をもとに行うこととなりますが、令和 6 年度の試行検証事業のなかで本問題が選定された理由として「記述式問題の中でも最も採点が困難と考えられる」という点が重視されたことに留意する必要があります。

採点が困難な問題に対して自動採点精度の測定や作業時間の推計を行っていることから、本項で得られた値は、記述式問題のなかでも比較的上限に近い範囲と考えられ、実際の調査では多様な難易度・文量の問題が出題されることに注意して、作業時間の推計を実施する必要があると考えられます。

3 本調査研究内テーマ B (CBT での調査実施等に関する試行・検証) の出題問題を用いた採点時間の推計

本試行検証事業では、テーマ B (CBT での調査実施等に関する試行検証) において出題された問題のうち、採点時間の推計に必要な範囲に対して採点作業を実施しました。

本項ではテーマ B 出題問題に対する採点作業結果をもとに、令和 9 年度以降の全面 CBT 化に向けて問題種別ごとに必要な作業工数の推計や、採点工程の改良案についてとりあげます。

3.1 テーマ B で出題された問題の概要

テーマ B の実施事業者より提供を受けた、テーマ B の中で出題・採点された問題の内訳は下の表のとおりです。

教科	選択式	短答式	記述式
小学校国語	31	0	0
小学校算数	11	15	0
中学校国語	25	2	0
中学校数学	14	11	0

本年度テーマ B における作問の都合上、小中学校の国語・算数数学では記述式問題が出題されなかったため、本項で行う国語・算数数学に対する工数推計では、選択式・短答式問題を対象とした場合の、CBT 実施時の効率化の程度を測定、分析しています。

3.2 作業時間の計測方法と留意点

項番3の推計にあたって行った作業時間の計測方法と、テーマB出題問題を用いた推計を100万人規模の調査に対して適用する際の留意点を本項で示します。

項目	内容
採点作業時間に含まれる作業	次の2点の作業で発生した時間を作業者が記録し、これをもって「結果提供可能な品質で採点結果を与える」までに必要な工数と扱った。 ・異なる2名による採点作業時間（独立2回採点に必要な時間） ・独立2回採点の結果によらず、全ての採点結果を確認し、不一致や誤採点が見られた場合に修正するための作業時間 なお、出題問題の性質とテーマBで収集された解答件数の都合上、全ての作業を人間によっておこなっている。
「100万件分の作業時間」の推計方法	テーマBで出題、収集された解答件数に対して生じた作業時間を、同じ比率で100万件分に広げた場合の時間で推計している。 ※1,000件の採点に対し記録された時間であれば1,000倍している。 また、作業時間は無解答・未提出を除いた、文字入力があった解答に対して行われた部分に対してのもので、実際の本体調査では問題の特性による無解答率の差などを考慮する必要がある。
作業者の能力	今回の採点作業に従事した者は全員が採点作業の従事経験を有しており、一般に行われる採点業務においては、上位採点者と扱われる能力を持っていたと推定される。
採点作業環境	今回の作業では、表形式のデータベースで管理された、サマリデータに対して作業を実施しているため、100万人規模で実施される際の採点システムとはインターフェース等が異なる可能性がある。

3.3 選択式問題の採点時間推計

選択式問題の特徴として、解答として入力される内容が、指定された選択肢に限定されるという特徴があります。また、各類型への分類も、原則として選ばれた選択肢やその組み合わせによってのみ定まるという特徴を持っています。

つまり、通常出題されるような選択式問題であれば、受検者数がどれだけ増えても、解答の種類の上限は選択肢や取りうる組み合わせの種類数のままで変わらず、受検者数によらず同一の採点手法によって採点作業が完了するとみなすことができます。

今回の試行検証事業でも、選択式問題は概ね次のような手順で採点を完了していました。

作業工程	作業内容
正誤処理関数の作成	選択肢番号（選択肢の組み合わせ）に対応して正誤や類型を出力する表計算関数を作成する
関数の貼り付け・実行	解答の一覧に対して、表計算関数を貼り付けて実行する
反映内容の確認	反映された採点結果・使用した関数が正しかったかどうか確認する

今回の採点作業では、選択式問題 1 問あたり最大 30 分程度で作業を完了していましたが、100 万人規模で出題する場合であっても、解答として入力される内容が選択肢の数に限られるのであれば、今回の試行検証と同様に、正誤処理を判定する関数を実行し、データベースを検索して結果を確認すれば採点作業は完了できると考えられます。

特に 100 万人規模で出題する際には、万が一の誤りがないように採点結果の入力状態を確認する時間が追加で発生すること考えられますが、これを考慮しても、項番 2.6 で推計した記述式問題に対する作業時間や、次の項で示す短答式問題に対する作業時間と比較して、無視できる程度に短い時間で採点作業を完了できるとみなして差し支えがないと考えることができます。

3.4 短答式問題の採点時間推計

本項では、短答式問題に対する作業時間をもとに、100 万人規模で短答式問題を出題した際に必要な作業時間を推計します。

なお、本試行検証で使用した個々の問題内容や採点基準については非公開となるため、本項で説明する問題内容の例示は、過年度の本体調査で出題され、解説資料等で公開されている問題を参考に、本試行検証で集計したものと同等の工数積算が見込まれる問題内容を、説明のために取り上げています。

推計のために必要な情報として、以降では表の中で、下の4項目について集計した結果を取り上げます。

項目	概要
解答種類数/受検者数の比率	句読点や記号などを含め完全一致する解答を1種類として扱い、解答種類数を受検者数で割った割合。 ここが10%であれば、解答者100万人に対して10万種類の解答文字列が存在し、10万件分の採点作業で全体が完了することになる。
CBT 解答 100 万件あたりの作業時間数	解答1件（1種類）あたりの作業時間を、上記の解答種類数と受検者数の比率で掛け算して求められた時間。
CBT 1 解答欄・解答 100 万件あたりの作業時間数	完答式問題のように、1小問の中に複数の解答欄がある場合、その組み合わせによって解答の種類数が増加し、また解答1件あたりを目視採点するための作業工数が増加する。 このため、実際の作業時間を小問内にある短答式解答欄の数で割って、解答欄1つあたりに要する時間を算出している。
PBT 解答 100 万件あたりの作業時間数	完全一致する解答を1種類としてまとめられない、PBTで目視採点した場合に100万件を処理するための推定時間。

これらの項目をもとに、各問題の特徴に由来した作業時間の推計結果について、次の①、②で示します。

①：想定解と同内容の文字列を、原則 1 種類（あるいは列挙できる程度）しか考えられない問題

例えば、特定の漢字の読みを問う問題や計算結果の値を求める問題などでは、正答と認められる文字列は、原則として 1 種類しか存在しません。

別解がある場合や類型をより細かく分類する場合でも、許容される解答文字列は事前に列挙できる程度にとどまるため、このような問題に対する採点では選択式問題と同様に、想定正答と一致する文字列に対しては正答（または所定の類型）へ、一致しない文字列に対しては誤答へ分類するの
が、採点作業の基本方針になると考えられます。

ただし選択式問題と異なる点として、短答式問題の形式上は、テキストボックス等への自由記述が認められていることから、全半角の違いやスペース等記号の有無などの表記ゆれに由来して、想定文字列以外の入力であっても、想定解と同一内容が解答されている可能性があります。

このため、機械的に分類した中に誤分類が発生していないかを確認する工程が必要になると考えられ、この作業を 100 万人分の解答群に対して行うための工数を推計する必要があります。

今回の試行検証事業で、想定正答として認められる入力が 1 種類に定まっていた問題を対象にした、各項目の集計結果は下の表のとおりです。

	平均値	中央値	最大値
解答種類数/受検者数の比率	14.55%	13.52%	27.08%
CBT 解答 100 万件あたりの作業時間数	1046 時間	903 時間	3108 時間
CBT 1 解答欄・解答 100 万件あたりの作業時間数	770 時間	672 時間	1415 時間
PBT 解答 100 万件あたりの作業時間数	6,236 時間	2,361 時間	11,111 時間

表の情報のうち、100 万件あたりの作業時間数が 3,108 時間と推計された問題は、解答欄 4 つで答える完答式問題で、これによって全体で解答文字列の種類や判断時間が増加したことが原因と考えられました。この点でも作業時間を推計するに当たっては、単に 1 問あたりの作業時間ではなく解答欄単位でみる方がより妥当性が高くなると考えることができます。

あらためて 1 解答欄あたりの作業時間数に注目すると、平均として約 800 時間程度と推計されました。各問題で作業時間の振れ幅はあるものの、このような想定解以外を認めない問題は調査全体では多数出題されることから、平均値と問題数（解答欄数）の掛け算によって、おおよその工数が近似的に推計可能と考えることができます。

また、これらの問題を PBT で出題しようとした場合と比較すると、1 解答あたりの平均で 5,000 時間以上短く完了できることが期待でき、この差は同一の解答入力を 1 つにまとめて採点できる、CBT 調査の利点によって得られた効果と解釈できます。

②：想定解と同一内容の文字列を、いくらでも考えることができる問題

例えば、短答式問題であっても文章による説明を求める問題などでは、想定正答以外の文字列であっても正答となることが十分に考えられ、記述式問題と同様に書かれた内容を読み取って判断する必要があります。

また、算数数学では、文字式や分数などによる解答を求める場合、式を組み換えたり違う数値を使って表現したりすることで、同一内容とみなせる解答をいくらでも生み出せる場合があります。

このように、事前想定にない許容解が多数存在できる問題の場合には、他よりも解答の種類が多くなると考えられるほか、各解答がどの類型に属するのか個別に確認していく必要性が高くなると考えられ、工数の推計にあたっては、①で述べたような問題とは区別する必要があります。

本試行検証で出題した問題のうち、想定正答以外の許容入力が多い問題を対象にした、各項目の集計結果は次の表のとおりです。

	平均値	中央値	最大値
解答種類数/受検者数の比率	30.73%	26.91%	59.10%
CBT 解答 100 万件あたりの作業時間数	3,373 時間	2,915 時間	5,910 時間
CBT 1 解答欄・解答 100 万件あたりの作業時間数	2,782 時間	2,915 時間	4,097 時間
PBT 解答 100 万件あたりの作業時間数	10,167 時間	10,833 時間	14,722 時間

ここでも 1 解答欄あたりの平均値でみると、自由な解答が可能な短答式問題に対する作業時間として、1 解答欄あたり約 2,800 時間という推計値が得られます。

自由記述が認められる分、受検者数に対する解答の種類は増加しており、これが作業時間数を増加させる要因となっていますが、PBT で出題した場合の作業時間としては 1 解答欄あたり 10,000 時間を超える推計値となり、同一の文字列に対して一括で採点できる CBT ならではの効果はこちらでも確認できます。

このように、項番 2.6 で推計された記述式問題に対する採点作業時間と比べれば、短答式問題であれば難しい内容であっても比較的短い時間で採点を終えることが確認されました。

しかしながら調査全体の作業工数推計にあたっては、児童生徒 1 人に対して何問の短答式問題が出題されるかを考慮して積算する必要があり、調査問題を設計するに際しても、1 問ごとに作業時間が積み上げられていく点にあらためての留意が必要です。

4 令和9年度の全面 CBT 調査における採点工数の推計

ここまでの推計結果をもとに、令和9年度に全面 CBT 化となる予定の、小学校国語・算数、中学校国語・数学のそれぞれで、採点作業に必要な全体の作業時間について推計を実施します。

なお、現時点で令和9年度調査のなかで出題される問題は不明であるため、ここで示された採点工数は、今回の分析に加え、過去の本体調査・経年調査から得られる公開情報から推計されたものである点に、強く留意が必要であることをはじめに示します。

4.1 工数推計の前提条件

本項で採点工数として示す作業時間は、次の表に示す前提条件が含まれています。実際の令和9年度以降の調査実施に際しては、出題内容によって推計条件が大幅に補正される可能性がある点に留意が必要です。

項目	推計結果を解釈する際の留意点
推計対象の工程	本項や本調査研究内で推計する作業時間は、自動採点を併用して独立2回採点を実施し、各解答に類型を付与する部分の時間として示している。 実際の調査のなかでは、採点基準・作業マニュアルの作成、出題問題の CBT システムへの搭載、採点者への研修・品質管理などの工数を別途積み立てる必要がある。
試行検証事業の作業者	令和6年度試行検証事業を含み、本試行検証で採点作業に関わった作業者は、ほとんど全員が上位採点者以上の能力を有した者となっている。 実際には、多数の一般採点者によって採点作業が実施されることから、速度や品質の面で、今回推計の根拠とした数値と差が生じることが予想される。

項目	推計結果を解釈する際の留意点
出題問題の構成	令和9年度以降の問題構成や内容は現時点で不明であるため、過年度の全国学力・学習状況調査の出題情報から、合理的に予想できる問題構成で出題された場合の工数を推計している。 特に CBT 調査にあたっては、1人あたりの問題構成や出題問題数が PBT から大きく変化する可能性があり、どの問題種別がどれだけ出題されるかによって、作業時間の推計は大きく変化する。
出題問題の内容	過年度 PBT で出題されていた問題の中には、現時点では CBT で出題困難な問題が含まれる場合がある。逆に、今回の推計は全てテキスト入力によって解答される問題群に対して行われたものなので、画像などのコンテンツを操作して解くような CBT ならではの問題であったり、国語・算数数学でも口述式による問題が出題されたりする場合などでは、工数の推計が困難となる。

4.2 過年度調査の構成を参考とした採点工数の推計

まず、作業時間の概算をより正確に推計するためには、「①：1人あたり何問出題されるか」、「②：問題の採点難易度はどれほどか」、「③：何人が受検するか」の3つの要素が必要です。

これらを掛けあわせることによって全体の作業時間を推計することが一定程度可能となりますが、前項で触れたように、現時点では出題問題の構成や内容が未確定であるため、本試行検証で得られた結果と、その他の公開情報から推計を実施します。

ここで、全国学力・学習状況調査に準じた構成の CBT 調査として、令和6年度全国学力・学習状況調査（経年変化分析調査）のなかで CBT 調査が実施され、この結果の一部が実施結果報告書として公開されています。

本項ではこの実施結果報告書を一部利用し、令和9年度の全国学力・学習状況調査の本体調査部分で、どの程度の作業時間が見積もられることになるかを推計します。

4.2.1 出題問題数の推計

特に CBT 調査における出題問題数のための参考情報として、令和6年度の経年変化分析調査では、各教科1単位時間（小学校40分、中学校45分）で CBT 調査が実施されています。

実施結果報告書の中では、出題問題数の総計と IRT 分析のための分冊構成が示されており、これから1単位時間で出題が見込まれる問題数について推計を行います。

まず、令和6年度経年変化分析調査（CBT 調査）で出題された問題数は下の表のとおりでした。

校種・教科	選択式	短答式	記述式
小学校国語	18 問	15 問	6 問
小学校算数	10 問	29 問	13 問
中学校国語	27 問	17 問	8 問
中学校数学	26 問	26 問	13 問

経年変化分析調査では、これらの問題群を 13 個のユニットに分割し、各児童生徒への出題にあたっては、13 ユニット中 4 ユニットの搭載した分冊 1 つを割り当てています。

ここで、経年変化分析調査で用いられている IRT 分析の性質上、分冊内で同じ問題が出題されないようにするために、各ユニット間では同じ問題は重複していないと考えられることから、児童生徒 1 人あたりに出題される問題数の期待値は、全出題問題のうち 13 分の 4 と推定することができます。

このため、令和 9 年度の CBT 調査においても同様に 1 単位時間の中で各教科の出題がなされると仮定した場合の、1 人あたりの出題問題数の期待値は下の表のとおり推定できます。

1 人あたり出題問題数	選択式	短答式	記述式
小学校国語	5.54 問	4.62 問	1.85 問
小学校算数	3.08 問	8.92 問	4 問
中学校国語	8.31 問	5.23 問	2.46 問
中学校数学	8 問	8 問	4 問

なお、項番 3.3 で述べたように、選択式の問題数については総作業時間の観点ではほとんど無視できるため、以降の推計では短答式と記述式の部分に対し着目した検討を行います。

4.2.2 出題問題内容に対する作業時間の推計

工数の推計にあたっては、1問あたりどの程度の作業時間が見込まれるかが、推計結果に大きく影響を与えます。このため、本来推計を行うにあたっては、その問題内容や解答傾向の詳細な情報が必要です。

しかしながら、経年変化分析調査の公表資料を含め、CBT 調査全体での出題内容や解答傾向の詳細は明らかでないため、実際の CBT 調査を仮定したときにどのような内容の問題が出題されることになるかわからない状態で推計を行う必要があり、本項もこれを前提に記述を行います。

加えて、本試行検証の中で行った推計としても、特に記述式については、「最も採点が困難と思われる問題」に対しての作業時間となっており、この数値をそのまま適用すると、「最も採点困難な問題だけで構成された調査」に対して推計されることになってしまいます。

難易度の観点や児童生徒に対する負担の観点からも、全設問がそのような構成になるとは考えにくく、簡単なものから難しいものまで様々な種類の問題が出題された場合の「平均的な記述式問題の採点時間」を仮定して推計しなければなりません。

このため本項では、難易度や問題内容、解答傾向などの要素については、「採点が簡単なものから難しいものまで、全て均等に、平均的に出題された場合」をやむをえず仮定して、以降の推計を実施するものとします。

①：短答式 1 問あたりの作業時間推計

項番 3.4 のなかで、想定解と同一内容の文字列を多数考えられるかどうかで、それぞれ集計された作業時間の平均値が異なることを示しました。

このため、今回の推計では、想定解と同一内容の文字列を考えられない問題と考えられる問題が、それぞれ同じ程度出題される場合を考えます。また、短答式 1 小問内の解答欄の数についても 1 つとは限らないため、解答欄あたりでなく問題あたりの作業時間をもとにして平均値を考えると、 $(1,046 + 3,373) \div 2$ より、1 問あたりの平均作業時間を約 2,209 時間と仮定することができます。

②：記述式 1 問あたりの作業時間推計

先ほど触れたように、記述式問題の推計結果は「最も採点が困難な問題」を仮定したものでした。実際には簡単な問題もあれば難しい問題もあり、「平均的な記述式問題の作業時間」を推計するためには、「最も採点が簡単な記述式問題」を仮定し、その中間値を仮定する必要があります。

ここで、令和 6 年度報告書で述べられた小学校算数の問題に対する自動採点のなかで、選択肢の誤りによって解答類型が定まってしまう、詳細な内容判断を要する解答が、有効解答 3,764 件中 637 件にとどまった旨が示されています。

これをもとにすると、「最も採点が簡単な記述式問題」として想定されるものの 1 つとして、例えば選択式と記述式の完答を要求する問題で、内容の判断が困難だったとしても、選択肢の誤りなどで大部分を形式的な処理で完了できてしまう問題を想定することができます。

小学校算数も、採点時の内容判断が困難であることを理由に選定されたことを考慮すると、項番 2.6 で求めた国語の内容判断と同程度の時間がかかるところを、形式的な処理で完了できる割合を求めることで、記述式問題に対する作業時間の下限を下のように考えることができます。

項目	「最も採点が簡単な記述式問題」で内容判断が必要だった割合	項番 2.6 で求めた、採点が困難な記述式問題に対する作業時間	記述式問題に対する採点時間の下限推計
計算数値	16.92% ※3764 件中 637 件	16,981 時間	2,874 時間

また、項番 3.4 で示された、短答式問題のうち自由記述が認められる問題に対する 1 問あたりの推計結果の平均 3,373 時間や、1 解答欄あたりの推計である平均 2,782 時間と照らしても、ここで求めた記述式問題に対する作業時間の下限は「自由記述問題であっても、類型の設定方法などによって、最良の場合で短答式と同程度の作業で済む場合がある」という評価を近似的に与えることができます。

このうえで、上限から下限まで幅広く出題された場合の記述式問題 1 問あたりの平均作業時間として、 $(16,981 + 2,874) \div 2 = \text{約 } 9,928 \text{ 時間}$ という推計値を考えることができます。

4.2.3 受検人数・有効解答件数の推計

ここまで各出題問題に対して推計した作業時間は、文字列の入力があった有効解答の 100 万件分に対するものであるため、実際には令和 9 年度本体調査に対する受検者数を推計した上で作業時間を求める必要があります。

加えて、実際の調査のなかで児童生徒が受検した結果としては、出題に対し無解答（文字入力なし）で提出されるものが存在し、この場合は明らかに採点作業時間を無視できるため、これを併せて考慮する必要があります。

まず、現時点で最新の令和 7 年度全国学力・学習状況調査において、後日実施を含めた受検人数は、実施概況にて示されています。

次に、令和 7 年度に実施された学校基本調査を用いると、令和 7 年度時点の小学 6 年生・中学 3 年生の人数と、令和 9 年度調査で受検が予定される小学 4 年生・中学 1 年生の人数比が求められるので、これによって下の表のとおり、令和 9 年度調査における受検人数の概算を推計することができます。

校種・教科	令和 7 年度 全国学調受検人数	令和 7 年度学校基本調査		令和 9 年度 全国学調推計受検人数
		小 6 ・ 中 3 児童生徒数	小 4 ・ 中 1 児童生徒数	
小学校国語	966,169 人	1,016,003 人	997,040 人	948,136 人
小学校算数	966,304 人			948,269 人
中学校国語	904,463 人	1,052,917 人	1,018,443 人	874,850 人
中学校数学	905,008 人			875,377 人

続いて、各教科の問題種別に対しての無解答率として、令和 6 年度経年変化分析調査のなかで、出題問題に対しての無解答率が示されています。この平均値を取ることで、全問題が均等に
出題された場合の、問題種別ごとの無解答率として下の表に示す割合が推計できます。

無解答率の平均	短答式	記述式
小学校国語	4.15%	8.98%
小学校算数	8.71%	16.27%
中学校国語	3.48%	11.08%
中学校数学	7.15%	18.32%

これにより、各教科・問題種別ごとに、出題 1 問あたり文字列入力がある解答数として、下の表の通り推計することができます。

有効解答件数の推計	短答式	記述式
小学校国語	908,788 件	862,993 件
小学校算数	865,675 件	793,986 件
中学校国語	844,405 件	777,917 件
中学校数学	812,788 件	715,008 件

4.2.4 総作業時間の概算と評価

項番 4.2.1 から項番 4.2.3 で示した情報を掛けあわせることで、総作業時間の概算を推計することができます。

問題種別ごとの推計計算と、最終的に求められた総作業時間の概算結果は下の表のとおりです。

①：短答式問題の作業時間概算

校種	推計問題数	有効解答 100 万 件あたり作業時間	推計解答件数	合計作業時間 の推計
小学校国語	4.62 問	1,776 時間	908,788 件	7,457 時間
小学校算数	8.92 問		865,675 件	13,714 時間
中学校国語	5.23 問		844,405 件	7,843 時間
中学校数学	8 問		812,788 件	11,548 時間

②：記述式問題の作業時間概算

校種	推計問題数	有効解答 100 万件 あたり作業時間	推計解答件数	合計作業時間の 推計
小学校国語	1.85 問	9,928 時間	908,788 件	16,692 時間
小学校算数	4 問		865,675 件	34,378 時間
中学校国語	2.46 問		844,405 件	20,623 時間
中学校数学	4 問		812,788 件	32,277 時間

③全体の作業時間推計結果

総作業時間推計	短答式	記述式	教科合計	調査合計
小学校国語	7,457 時間	16,692 時間	24,149 時間	72,241 時間
小学校算数	13,714 時間	34,378 時間	48,092 時間	
中学校国語	7,843 時間	20,623 時間	28,466 時間	72,291 時間
中学校数学	11,548 時間	32,277 時間	43,825 時間	

※参考：全ての問題で目視のみによる採点を仮定した場合の推計（計算過程略）

総作業時間推計	短答式	記述式	教科合計	調査合計
小学校国語	34,435 時間	21,648 時間	56,083 時間	162,478 時間
小学校算数	63,331 時間	43,064 時間	106,395 時間	
中学校国語	36,220 時間	25,948 時間	62,168 時間	154,278 時間
中学校数学	53,329 時間	38,781 時間	92,110 時間	

※目視の場合の記述式問題の 100 万件あたり採点時間概算は、令和 6 年度試行検証で得られた 1 件あたり 83.5 秒より、概算 23,194 時間。その他出題問題数や「平均的な時間」の算出方法などは、上記の項番 4.2.1～4.2.3 と同様。

表のとおり、全面 CBT 化による調査実施の際に推計される作業時間の概算値として、小中学校とも、約 72,000 時間という結果が得られました。

特に PBT で出題される際に想定される、全てを目視による採点を行った場合の概算と比較すると、どの教科でも多数出題される短答式問題に対して、同一の解答を一度に採点できることによる効果が非常に大きく、また、記述式についても、自動採点プログラムと併せて形式的な採点工程の導入による効果が現れていることが推定できます。

一方で、全面 CBT 化を仮定した場合でも、人間による作業時間として約 72,000 時間という概算が得られたことに注目すると、特に実際の採点作業に際しては、この作業時間によって所定の採点期間内で完了可能かどうか、注意して検討する必要があります。

令和 7 年度全国学調では、CBT 調査に対する採点期間の仕様として、PBT では調査日から 6 週間という要件と比較して CBT では 4 週間以内での採点完了が求められており、これと同水準の期間短縮が令和 9 年度調査においても求められる場合、1 日あたりの必要作業時間として、単純計算でも、 $72,000 \div 28 = \text{約 } 2,570$ 時間という概算が与えられます。

特に大規模な採点においては、これらの作業時間を一般の採点者によって賄う必要があると考えられ、項番 4.1 で触れたような、採点者を熟練させるための研修や、採点品質の管理工程を含めた作業計画が、採点期間内に収まるかどうかを検討することが強く求められると予想されます。

加えて、今回推計した作業時間は、問題数や出題内容について複数の仮定を置いている点にも同程度に強く留意が必要です。

最も極端な場合を考えると、今回の仮定から 1 問記述式問題の出題を増やした場合、最大で約 17,000 時間の作業時間が加算されると推計されてしまうことから、今回の推計に使用した令和 6 年度経年変化分析調査からどの程度の差分が見込まれるのかも考慮しながら、調査の設計を進める必要があると考えられます。

5 まとめ

本報告では、令和9年度全国学力・学習状況調査（悉皆調査）において採点作業に必要な時間を推計するため、各種問題に対する採点時間を推計すると同時に、現時点で想定できる採点工程の効率化手法について測定・検証を行いました。

特に採点工程の効率化に対して検証された内容は、CBT 調査による長所として、児童生徒の解答全体を大規模データとして扱える特徴を活かすようなものとなりました。

一方で、令和9年度の調査全体での採点作業工数を推計するにあたっては、出題問題の規模や内容に対して、令和6年度全国学力・学習状況調査（経年変化分析調査）をもとに、複数の仮定を置いたうえで求めることになりました。

1問あたりの採点作業について効率化できることはここまでの試行検証事業から示されましたが、調査全体でどの程度の時間が必要となるのかの検討に際しては、実際に出題される問題数と内容に依存するという要因は避けて通れません。

加えて、調査の CBT 化にあたり期待される事柄として、採点作業の効率化と同時に、採点完了までの時間を短縮することも併せて求められることとなります。従来と比べて期間を短縮したとき、全体の作業規模がどう変化するかという観点でも、実施にあたっては留意が必要です。

自動採点に関わる技術については、社会状況としては昨年度から引き続いて向上が見られる分野です。他方で、全国学力・学習状況調査という事業全体で見たときに、期間内で採点作業を完了させるためには、採点技術の効率化だけでなく、出題問題の構成などを含めた、複数の要因について引き続きの検討が求められると考えられます。

以上