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
A.　CBT記述式答案の採点に関する試行・検証

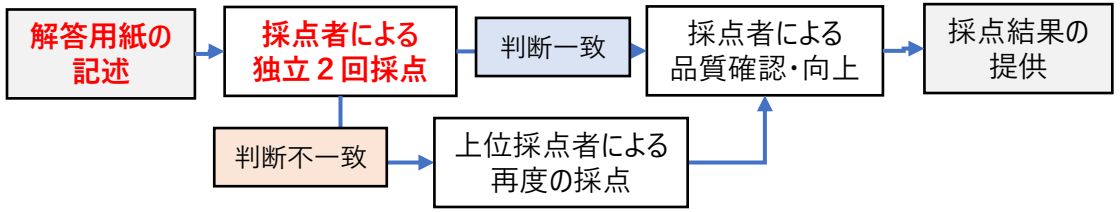
最終報告書（概要版）

令和７年３月

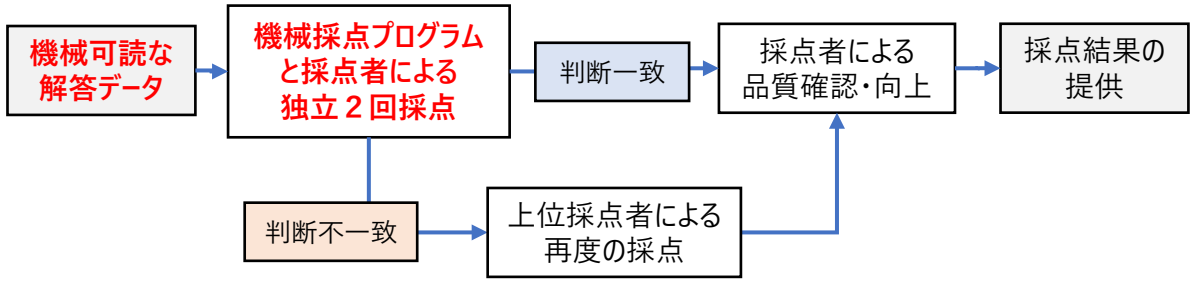
本事業の目的

全国学力・学習状況調査の段階的なCBT化にともない、機械可読な解答データを用いた機械採点プログラムの導入による効率化の程度を検証する。

PBTの採点プロセス概略：
答案用紙を人間の採点者２名が目視により採点し、採点結果が一致しない場合は上位採点者が採点する。
採点結果に対しては再度人間が品質確認を実施する。

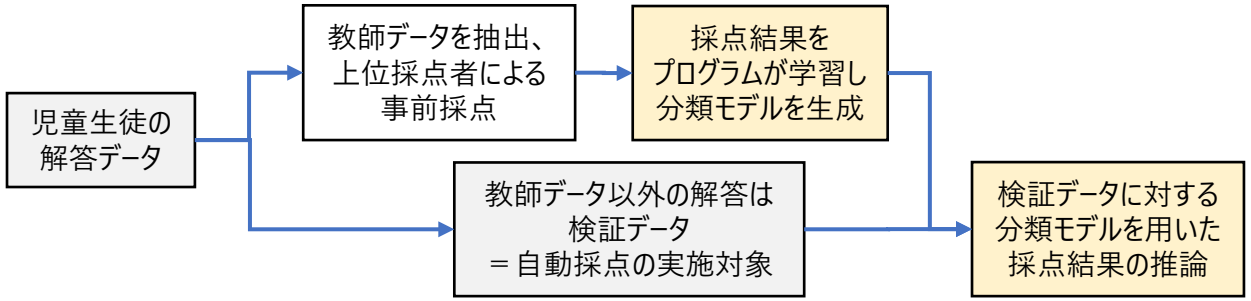


CBTで期待される採点プロセス概略：
解答データに対し機械採点プログラムによる自動採点を実施するとともに、人間の採点者も従来どおり採点作業を受け持ち、これまでと同様、独立２回採点のプロセスによって採点する。
採点結果の確認もPBTと同様に人間が行い、最終的な品質を担保する。



機械採点プログラムの概要

収集された児童生徒の解答の一部を習熟度の高い上位採点者が事前に採点し、
採点結果を教師データとして機械採点プログラムが学習して、問題ごとに分類モデルを生成する。
残りの解答データに対しては分類モデルで推論を実施して、自動採点の結果として出力する。



検証の概要

「令和６年度全国学力・学習状況調査 経年変化分析調査」記述式問題や「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テーマB　CBTでの調査実施等に関する試行・検証」の実施問題について自動採点を実施する。

検証結果をもとに、機械採点の精度に影響を与えると考えられる要因の整理や、全国学力・学習状況調査での自動採点の活用の可能性やスケジュールの推計を行う。

※調査の性質上、個々の問題内容、採点基準等は非公開である。

事業成果：国語と数学では、機械採点導入の効果が一定程度確認できた。

記述式問題：採点精度や効率化程度の測定に加え、採点精度に影響を与える要因について質的分析を行った。

今回用いた機械採点プログラムにおいて精度としては向上の余地を残すものの、従来人間が実施していた採点作業プロセスの一部を適切に置き換えることで、採点業務全体としては効率化が期待できることが推計された。
併せて、教師データ件数と類型一致率の関係から、機械採点に必要な教師データ作成のための工数を充分に確保し、多数の教師データを使用することの重要性も示唆された。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。

小学校国語の長文記述問題に対する、
機械採点による類型一致率の測定結果

教師データ件数	最低値	最高値
500件	59.23%	76.57%
1,000件	77.01%	81.28%

※教師データはランダムサンプリングで抽出したため、教師データの内容により測定結果が変化する。
数値は今回用いたプログラム・問題・解答に対する結果で、普遍的なものではない。

また、機械採点の精度に影響を与えると考えられる、問題内容、解答内容、採点基準の特徴について、下に示すとおり整理した。

正答の条件・類型の数	解答（正答）の多様性	1 文字の違いに由来する採点基準	形式的に定まる採点基準
推論しなければならない正答の条件や、分類すべき類型の種類が増えるほど、条件全てを正しく推論できる確率は低下すると推測される。	問いに対して許容される表現の幅が広い問題では、採点難易度が上昇する傾向が見られた。	数式や英単語のように、1 文字の差によって正誤や類型に影響する場合、1 文字の差による意味内容の変わり方を正確に検知する必要があるため、採点難易度が上昇する傾向が見られた。	選択肢のように形式的に一意に定まる採点基準が存在する場合には、意味内容を解釈する手法だけではなく、関数による処理のような手法を組み合わせることで、より正確に分類が可能である。

短答式問題：自由記述を認める問題であっても、特定の頻出解答が多数を占める傾向が確認された。

特に知識・技能を問う性質の問題では、児童生徒が解答する文字列は、設問に沿った一定の範囲に集中すると推測される。
記述内容を解釈するプログラム以外にも、同一の解答文字列に対し同時に採点を行う仕組みを導入することで、必要な採点件数の圧縮が期待できる。

今後の課題：英語記述式問題（特に長文で記述解答を求める問題）に対する自動採点の活用手法

場面設定・活動に由来する、内容表現の幅広さ

特に英語の長文記述問題では、問題の場面設定に由来して、他教科の記述式問題と比べ許容される解答の範囲が広くなる場合があり、児童生徒個人の経験や主観を含む解答に対し推論を実施する必要性によって、国語や数学の長文記述問題よりも採点難易度が高くなったと推測された。

教科特性に由来する、1 文字単位の判断要求

特に英語では文法の正しさが評価対象となる特性上、同じ意味内容でも、誤字や文法の誤りが他教科よりも類型や正誤に反映されやすい傾向がある。
このため、記述の意味内容を解釈する機能だけではなく、語句単位の探索などを機械採点のプロセス内に適切に組み合わせることの必要性が示唆された。悉皆調査への導入にあたり、人間による採点との適切な役割分担についての検討が必要である。