

論述採点の正確さと所要時間に関する研究

○野澤雄樹* 堂下雄輝* 島田研児**

* ベネッセ教育総合教育研究所 ** 株式会社ベネッセ i-キャリア

問題と目的

現行の学習指導要領では、知識・技能の習得だけでなく、それらを活用して課題を解決する能力（思考力・判断力・表現力を中心とした汎用的な能力）の育成が重視されている（文部科学省，2009）。この動きに合わせて、教育の成果を測定するためのテストも、汎用的な能力の測定に対応することが求められている。思考力・判断力・表現力などを測定するテストでは、より複雑な認知能力を測定する必要性から、記述・論述式の項目が大きな割合を占めるようになると考えられる。

記述・論述式の項目を多く含むテストを実施する場合、採点の管理が問題になる。本来なら、採点者を同じ会場に集め、熟達した採点リーダーのもとで、正確さや進行状況をモニタリングしながら、採点を進めることが望ましい。しかし、よほど重要度が高いテストでない限り、厳格な採点管理体制を維持することは不可能である。採点管理に多くの費用をかけられないテストでは、個々の採点者のパフォーマンスに依存する部分が大きくなると思われる。

このような状況を踏まえて、本研究では、採点者の自主性に任せて採点してもらった場合に、採点の正確さや所要時間がどのように変化するのか調べた。採点者の行動に関する情報を収集し、費用が少ない採点管理方法を考える上での基礎資料にすることが目的である。

データの収集

論述課題：大学生の「資料活用力」を測定するために開発された論述課題を使用した。この課題では、主張すべき内容があらかじめ決められている。受検者は、与えられた複数の資料からこの主張を支える根拠となる情報を抽出し、文章にまとめることが求められる。開発者が意図した根拠が挙げられているか、根拠と主張を結ぶ論拠が明確であるかが評価基準となる。解答時間は50分で、字数制限は特に設けられていないが、600字程度を想定している。

受検者：調査は2013年の7月下旬から8月上旬にかけて行われた。首都圏の大学に通う学生を都内のテストセンターに集め、上記の課題をコンピュータで実施し、205枚の答案を得た。

採点者：テストの採点を請け負う業者の社員2名と、この業者に登録している2名の採点者に採点を依頼した（男性が1名、女性が3名）。論述式テストの採点経験があったのは1名だけで、残りの3名は短答式テストの採点経験があるだけだった。

採点手続き：採点は、各採点者が各答案に対して、7つある評価ポイントのそれぞれについて、2段階あるいは3段階で評価値を付ける形で行われた。また、各採点者に対し、それぞれの答案について、採点の自信度と答案の読みやすさをそれぞれ2段階（自信あり、自信なし）と3段階（読

みやすい、読みにくい、どちらでもない) で回答するように求めた。答案の並び順はすべての採点者で同じにした。採点が終了した答案を見返すことや、評価値を後から変更することは禁止した。一方で、採点ルールの詳細を記したマニュアルを適宜参照することは許可した。採点は、2015年の2月下旬から3月上旬の間に行うよう求めた。

所要時間の記録：各採点者に対し、指定した80枚の答案(6-25, 66-85, 126-145, 186-205)について、採点に要した時間を秒単位で記録するように求めた。これらの答案を採点する際には、採点に直接関係がない作業はしないように教示した。採点マニュアルを参照した場合は、その時間を所要時間に含めるように教示した。

事前トレーニング：採点に先立って4人の採点者に同じ会場に集ってもらい、各評価ポイントの各評価値がどのような答案に対して適用されるのか説明した。その後、練習用の答案を使って実際に採点してもらい、基本的な採点ルールが理解されていることを確認した。事前トレーニングに要した時間は約3時間だった。

データの分析

開発者が定めた変換式に基づき、各採点者から得られた評価値データを1~5点のスコアに変換した。これらのスコアを各答案に付けられた基準スコア(開発者が同じ採点ルールで各答案に付けた評価値を変換して得られたスコア)と比較した。採点者のスコアが基準スコアと一致した場合には1、一致しなかった場合には0でコード化し、「正確度」変数とした。

採点の正確さの分析では、採点者ごとに「正確度」を応答変数とした一般化加法モデル(generalized additive models; Wood, 2006)による分析を行った。採点の所要時間の分析では、採点者ごとに「所要時間」を応答変数とした一般化線形モデルによる分析を行った。どちらの分析においても、表1に示した説明変数のうち「採点順」をまず投入し、それ以外は優先順位とAICを考慮しながらにモデルに追加した。

表 1. 説明変数のリスト

変数名 (優先順位)	取り扱い	備考
採点順 (1)	連続	所要時間の分析時には、最初の答案が0になるように1を引いた
文字数 (2)	連続	基準を600文字にするため、実際の文字数から600を引いた
基準スコア (2)	カテゴリカル	1~5点 5点を基準カテゴリーとした
採点の自信度 (3)	カテゴリカル	1 = 自信あり, 0 = 自信なし
答案の読みやすさ (3)	連続変数	1 = 読みやすい, 0 = どちらでもない, -1 = 読みにくい

結果

採点の正確さの分析：「正確度」と「採点順」を採点者ごとにプロットし、「正確度」が1になる確率(基準スコアとの一致率)を一般化加法モデルで推定した結果を図1に示した(図中の点線は95パーセント信頼区間である)。この図から、採点順の効果は採点者ごとに異なることが読み取れる(ただし、採点順の効果が5パーセント水準で有意だったのは採点者3のみである)。

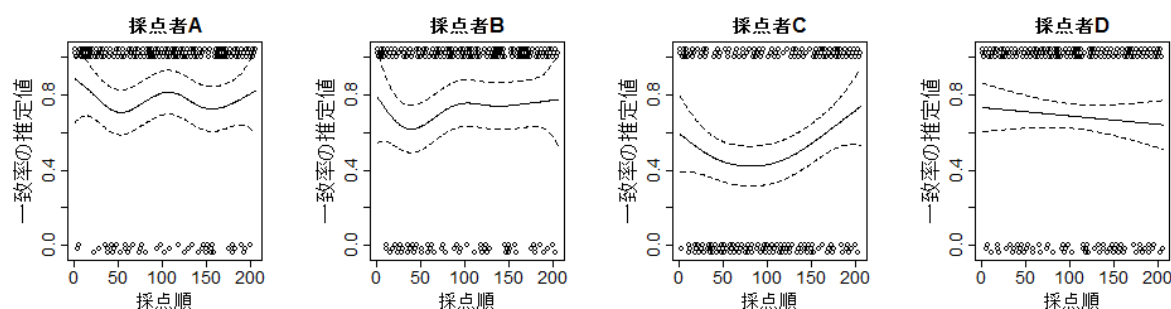


図 1. 基準スコアとの一一致率の推定値

計算は R の `mgcv` パッケージのバージョン 1.8-7 (Wood, 2015) に含まれる `gam` 関数で行った。

リンク関数をロジット、分布族を二項分布にした以外は、デフォルトの設定を使用した。

4 人の採点者に共通して、採点開始時は一致率が高く、その後一時的に（採点者 D は一貫して）一致率が下がるという傾向が見られた。採点しやすい答案が最初の方に集まっている可能性を取り除くため、「文字数」と「基準スコア」を説明変数に加えた分析を行ったが、採点者 B の採点順の効果が右上がりの直線として推定されたことを除けば、採点順の効果の形状に目立った変化は見られなかった。このことから、採点者 B 以外では、答案の性質を考慮しても、採点開始時の一致率が高めであったといえる。

採点の所要時間の分析：「所要時間」と「採点順」を採点者ごとにプロットしたものを図 2 に示した。図 2 から、所要時間の条件付き期待値は線形モデルで表現可能であると考えられる一方で、条件付き分散は条件付き期待値が小さくなるにつれて小さくなる傾向がみられた。そのため、一般化線形モデルを用い、条件付き分布をガンマ分布とした回帰分析を行った。図 2 に示されている直線は、説明変数として「採点順」のみを加えたときのものである（点線は 95 パーセントの信頼区間である）。

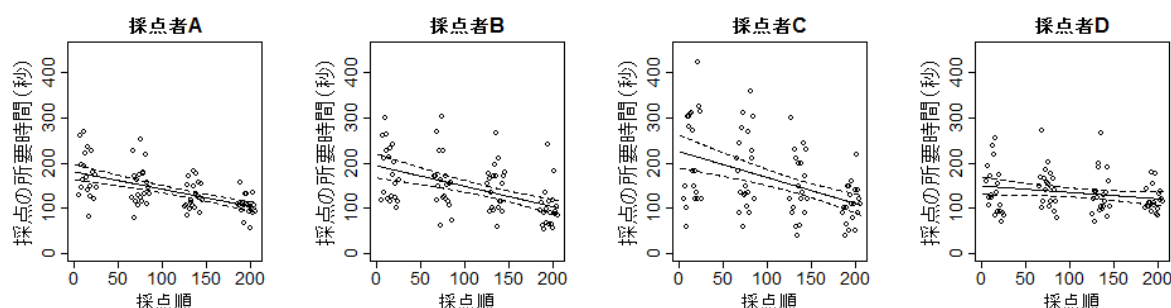


図 2. 採点順と所要時間のプロット

説明変数を追加して行った回帰分析の結果を表 2 に示した。採点順に対する係数はすべての採点者でマイナスであることから、採点が進むにつれて所要時間が減少するという傾向は、採点者間で一貫していることがわかる。逆に、文字数に対する係数はすべての採点者でプラスであることから、文字数が増えると所要時間が増えるという傾向も一貫していることがわかる。基準スコ

アが 5 点の答案に比べて、2～4 点の答案では所要時間が増える傾向がみられた。また、採点者が自信を持って採点した答案では、所要時間が 26 秒から 45 秒ほど短いという結果が得られた。採点者 A と B では、読みやすいと感じた答案のほうが所要時間を短い傾向が見られたが、採点者 C と D ではこの変数の効果が小さかったため、分析には含まれなかった。

表 2. 採点の所要時間の回帰分析結果（係数と標準誤差のみ）

説明変数		採点者 A		採点者 B		採点者 C		採点者 D	
		係数	標準誤差	係数	標準誤差	係数	標準誤差	係数	標準誤差
切片		201.190*	13.483	190.376*	14.111	218.321*	18.353	137.915*	10.582
採点順		-0.308*	0.048	-0.362*	0.055	-0.586*	0.088	-0.092	0.052
文字数		0.059*	0.029	0.150*	0.036	0.042	0.050	0.090*	0.035
基準スコア	4	20.999	10.937	26.202	13.344	34.626	21.948	19.205	12.309
	3	11.665	9.671	37.142*	13.977	16.985	19.881	29.137*	10.755
	2	18.369	9.293	17.060	10.372	42.040*	19.323	26.924*	10.559
	1	-6.925	8.841	-2.807	10.573	-28.599	16.938	35.448*	10.709
自信度		-25.712*	7.830	-34.490*	10.567	-44.851*	13.294	-37.964*	8.044
読みやすさ		-27.162*	9.362	-12.309*	4.321	-	-	-	-

計算は R の glm 関数を用いて行った。

考察

本研究では、最初に採点トレーニングを行った以外は、採点者の自主性に任せて採点が行われた。採点マニュアルの参照は許可されていたため、採点の進行にともなって正確さが上がることも予想されたが、採点順の効果は採点者によってさまざまであった。一方、採点の所要時間は採点者間で一貫した傾向が見られた。採点が進むに連れて所要時間が減ることや、文字数が増えることで所要時間が増えること、出来の良い答案よりも中間の答案のほうが時間がかかることなどは、納得できる結果であるといえる。本研究の結果から、採点の効率はある程度自動的に上がっていくものの、採点の正確さは自動的にには上がらないため、採点ルールの定期的な確認を義務付けるなど、正確さを上げるための介入が必要であることが示唆された。

引用文献

文部科学省 (2009). 高等学校学習指導要領 文部科学省.

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Chapman & Hall/CRC.

Wood, S. N. (2015). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*. R package version 1.8-7.