日本テスト学会　第13回大会
関西大学　千里山キャンパス
2015年9月11日　一般セッション4

# 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**
* ベネッセ教育総合研究所
** 株式会社ベネッセ i-キャリア

# 背景

- 現在，知識・技能の習得だけでなく，思考力・判断力・表現力などを重視した教育への転換が進められている。
  - 学校教育法の改正，学習指導要領の改訂
  - 世界的な学力観の変化
- このような流れを受けて，今後，教育効果を測定するテストに，記述・論述式の問題が増えると予想される。
  - もちろん，思考力などを測定するのに記述・論述式テストが有効かについては議論の余地がある。

# 背景

- 記述・論述式の問題が増えると，採点の質保証が課題になる。
  - 採点の正確さを担保できるか？
  - 採点を期間内に終了させられるか？
- 一般的なテスト・プログラムが，厳格な採点管理システムを構築するのは難しい。
  - お金だけでなく，時間も膨大にかかるため。
  - 結果として，採点の正確さが，採点者個人の資質に依存するということが起こってくると考えられる。

# リサーチクエスチョン

- 採点基準が比較的明確な論述課題において，事前トレーニングだけを実施し，その後は採点者に任せた場合に，採点の正確さは，採点の進行とともにどのように変化するか？

- 採点の所要時間は，採点の進行とともにどのように変化するか？

# 論述データの収集

**課題**：「資料活用力」を測定するために開発された論述問題
（制限時間50分，字数制限なし，目安は600字）

**受検対象**：首都圏の大学に通う大学生

**テスト形式**：都内のテストセンターでCBT受検

**実施**：2013年7月下旬~8月上旬

# 課題の内容

ある証券会社の社員になったという設定で，もうすぐ定年退職を迎える人物から相談を受けて，手紙で返答するという課題。

相談内容は，退職金の一部を投資運用することで，将来の経済的な不安を解消できないか，というもの。

解決策として自社商品を薦めることになるが，その根拠となる情報を与えられた資料から抽出し，論理的に説明することが要求される。

# 採点項目

- 自社商品が問題を解決できることを示せている（2項目）
- 他社商品では問題を解決できないことを示せている（4項目）
- 致命的な資料の読み取り誤りがない（1項目）

- 採点者は，答案の具体的な記述に基づいて，7つの採点項目に2段階あるいは3段階で評価値を付ける。

- 各採点項目に対して入力された評価値に基づいて，1～5点のスコアが与えられる。

# 採点データの収集

**採点者**：採点を請け負う業者を通じ，4人に採点を依頼。論述式の採点経験があったのは1人のみ。

**採点順**：3時間ほどの事前トレーニングの後，205枚の答案を同じ順番で採点してもらった。

**所要時間**：一部の答案（80枚）について採点に要した時間を記録してもらった。

**その他**：すべての答案について，「採点の自信度」と「答案の読みやさ」を評価してもらった。

**採点期間**：2015年2月下旬～3月上旬

# 採点の正確さの数値化

- この問題の開発者チームが205枚の答案を採点し，各答案の「基準スコア」を算出した。

- 各採点者について，スコアが基準スコアと一致した場合は1，しなかった場合は0の値をとる「正確度」変数を作成した。

# 分析で使用する変数の一覧

| 応答変数 | 概要 |
|---|---|
| 正確度 | 基準スコアとの一致を表す二値変数 |
| 所要時間 | 採点に要した時間（秒） |

| 説明変数 | 概要 | 取り扱い | 優先順位 |
|---|---|---|---|
| 採点順 | 答案の採点順（-0 or -1） | 連続 | 1 |
| 文字数 | 各答案の文字数-600 | 連続 | 2 |
| 基準スコア | 開発者採点によるスコア<br>5が基準カテゴリー | カテゴリカル | 2 |
| 採点の自信度 | 1 = 自信あり<br>0 = 自信なし | カテゴリカル | 3 |
| 答案の読みやすさ | 1 = 読みやすい<br>0 = どちらでもない<br>-1 = 読みにくい | 連続 | 3 |

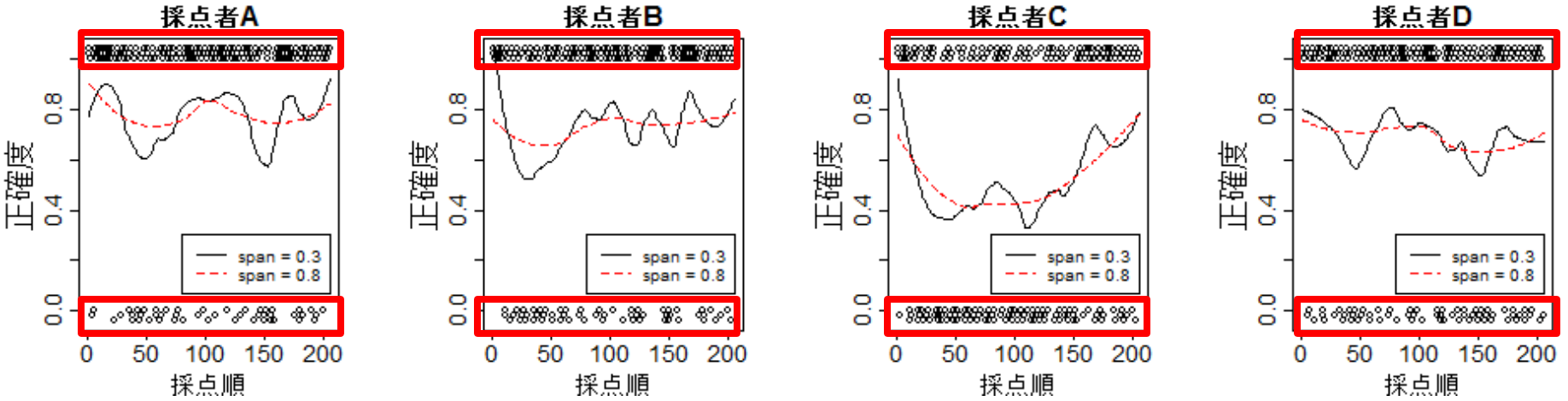
# 分析の各段階で投入する説明変数

段階1：「採点順」を投入。

段階2：「採点順」は必ず投入。「文字数」と「基準スコア」を別々あるいは同時に投入し，AICが最も小さかったモデルを選択。

段階3：段階2で選択された変数は必ず投入。「採点の自信度」と「答案の読みやすさ」を別々あるいは同時に投入し，AICが最も小さかったモデルを選択。

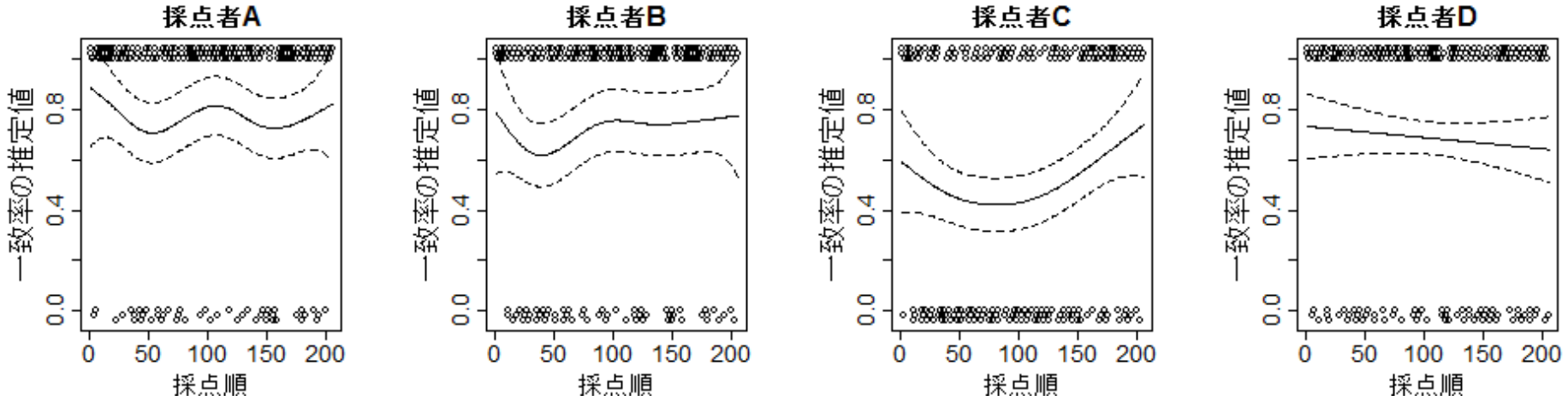
# 正確度のプロット




図中の線は loess を使って平滑化した結果。
平均一致率：**A = .77**, B = .72, **C = .52**, D = .69

- 採点者Aの一致率が総じて高く，採点者Cは低い。ただし，採点者Cは後半に一致率が上昇している。
- 通常のロジスティック回帰では採点順の効果をうまく表現できないため，一般化加法モデル（GAM）を使って分析する。[12]

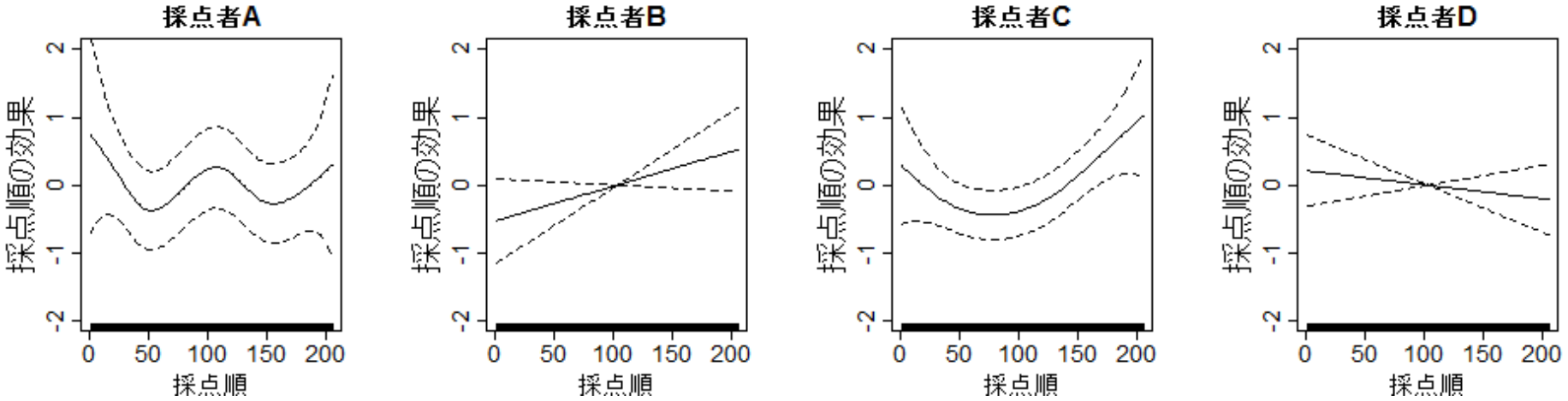
# 正確度のGAM分析（段階1）




実線はGAMを使って一致率を推定した結果。
破線は95%の信頼区間。

- 採点順のみ投入したGAMの結果。loess で span = 0.8 に指定した時の結果とかなり近い。
- 採点開始時の一致率が高く，その後下降する傾向があるように見えるが，誤差が大きいので，はっきりとは言えない。
- 採点順の効果が5%水準で有意だったのは採点者Cのみ。

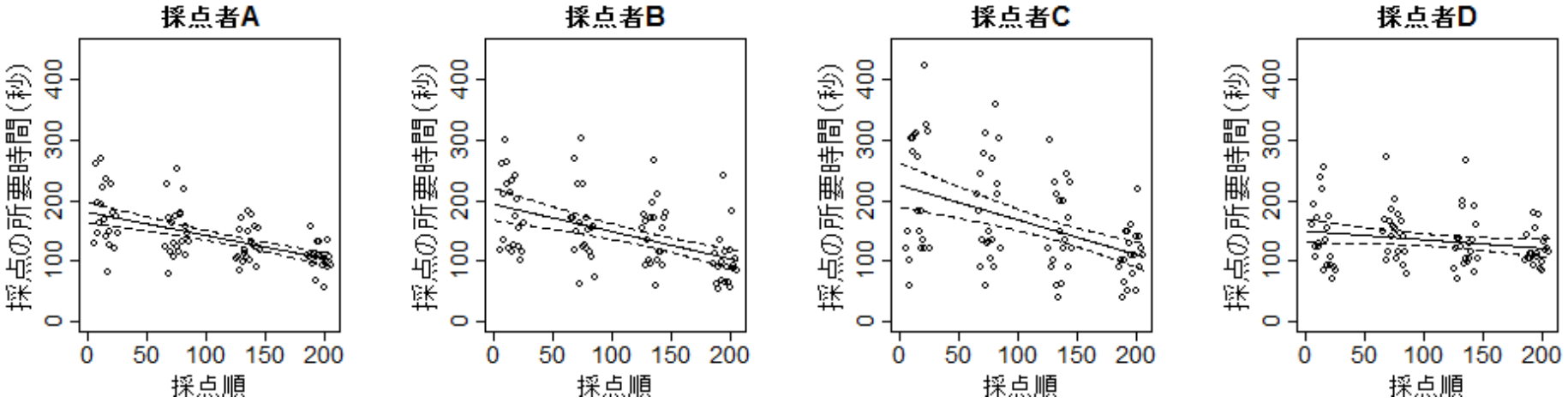
# 正確度のGAM分析（段階2）




実線は線形予測子内での採点順の効果。
破線は95%の信頼区間。

- 答案による影響を取り除くため，「文字数」と「基準スコア」を投入し，AICに基づいて取捨選択した。
  - 採点者BとCで「基準スコア」が追加された。
- 採点者Bの採点順の効果は理想的な形を示しているが，これは5%水準で有意ではなかった。

# 所要時間のプロット




実線はGLMを使って推定した回帰直線。
破線は95%の信頼区間。

- 所要時間の期待値は直線でうまく表現できそうだが，等分散を仮定するのは難しい。
  - 一般化線形モデル（GLM）を使い，リンク関数はそのままで，条件付き分布をガンマ分布に指定して分析を行った。

# 所要時間の分析結果（段階3の回帰係数）

| 説明変数 | | 採点者A | 採点者B | 採点者C | 採点者D |
|---|---|---|---|---|---|
| 切片 | | 201.190 | 190.376 | 218.321 | 137.915 |
| 採点順 | | -0.308 | -0.362 | -0.586 | -0.092 |
| 文字数 | | 0.059 | 0.150 | 0.042 | 0.090 |
| 基準スコア | 4 | 20.999 | 26.202 | 34.626 | 19.205 |
| | 3 | 11.665 | 37.142 | 16.985 | 29.137 |
| | 2 | 18.369 | 17.060 | 42.040 | 26.924 |
| | 1 | -6.925 | -2.807 | -28.599 | 35.448 |
| 採点の自信度 | | -25.712 | -34.490 | -44.851 | -37.964 |
| 答案の読みやすさ | | -27.162 | -12.309 | - | - |
| McFaddenの擬似的な$R^2$ | | .637 | .656 | .547 | .452 |

# 採点の正確さに関する考察

- 採点の進行とともに採点の正確さは変動するように見えるが，採点者間で一貫した傾向はつかめなかった。
  - 誤差が大きいため，採点者C以外は，実は正確さが一定であるという解釈も捨てきれない。
- 正確度が二値変数であること，採点者数が4人と少ないことの限界。
- 一方で，採点者Cが急回復した理由を調べることが，効率のよい採点者トレーニングに関して示唆を与えてくれる可能性がある。

# 採点の所要時間に関する考察

- 所要時間の分析結果は感覚的に理解しやすいものが多かった。
  - 採点が進むにつれて，所要時間が短くなっていくなど。
- 採点者間の回帰係数の差は，採点者の性質によってある程度説明できそうである。
  - 論述採点経験者の採点者Dは切片が小さいなど。
- 採点者数を増やし，採点者の属性情報を含めた階層モデル分析を行ってみたい。

ご清聴ありがとうございました