日本テスト学会第14回大会
一般研究セッション2 「ジェネリックスキル」
電気通信大学　第2会場（B102）
2016/09/09　9:00-10:20

論述力を測定する
混合フォーマットテスト開発の試み

野澤雄樹　伊藤素江　須永正巳　堂下雄輝　村田維沙
ベネッセ教育総合研究所

日本テスト学会第14回大会
電気通信大学
2016年9月9日

混合フォーマットテストとは？

- 多肢選択式と論述式のように，異なる形式の問題を組み合わせて作るテストのこと。
  - 英語ではmixed-format testと呼ばれる。
- 混合フォーマットテストを用いる理由として，例えば論述式の問題のみでテストを作ろうとすると，信頼性を高くすることが難しいということが挙げられる。
  - 論述式の問題は1問あたりの解答時間が長く，採点コストも大きいため，たくさん出題することができない。
  - その結果，受検者と問題の交互作用に起因する誤差分散を小さくすることが難しい。

混合フォーマットテストとは？

- 多肢選択式の問題は出題数を比較的多めにできるので，信頼性が高いテストを作成するのに効果的である。
  - とはいえ，すべての問題を多肢選択式にしてしまうと，論述式の問題でしか測定できない構成概念の側面を取りこぼしてしまう可能性がある。
- 多肢選択式と論述式の問題を組み合わせてテストを構成することで，お互いの弱点をカバーすることができる。
  - 本研究では踏み込まないが，多肢選択式の問題は等化を行う際の共通項目として使いやすいというメリットもある。

論述力の測定とその課題

- 本研究においては，論述力を以下のように定義する。

> 目的に応じてインプットした情報を考察し，考察した内容を論理的・効果的に言語で表現する能力。

  - この定義に従えば，論述力を論述式の問題で測定するのは当然のように思われる。

- しかし，現在開発中の論述式の問題は，1問あたり30分程度の解答時間が必要なので，テスト時間を長めに設定しても，3問程度しか出題できないという課題がある。

本研究の目的

- 論述式と多肢選択式（あるいは短答式）の問題を併用することで，論述力の測定を正確に行えるようにしたい。

- とはいえ，論述力は総合的な能力のため，組み合わせる問題もさまざまなものが考えられる。
  - 思考力，読解力，あるいは語彙力を測る問題？

- 本研究の目的は，どのような問題をどれぐらい併用したときに，論述力の測定の正確さが向上するのかを調べることである。

データの収集

- 首都圏の大学に通う学生を対象にして，2015年12月に都内のテストセンターで，以下の6つのテストをコンピューターで実施した。
  - 論述テスト（後述）
  - 批判的思考テスト（後述）
  - 汎用読解テスト（連続・非連続テキストから成るPISA型のテスト）
  - 教科型読解テスト（評論文の論展開・趣旨を問うテスト）
  - 短答式・語彙読解テスト（語の活用方法や論構成の理解を問うテスト）
  - 語彙テスト（芝 (1978) に基づいて作成されたテスト）

- 以下では，論述テスト以外のテストをまとめて「客観テスト」と呼ぶ。

テストの詳細

テスト名	形式	問題数	採点	得点	試験時間	受検者数
論述	論述	3	5段階評価	評定値の合計	110分	167
批判的思考	多肢選択	40	正誤	正答数	60分	500
汎用読解	多肢選択	25	正誤	正答数	40分	500
教科型読解	多肢選択	17	正誤	正答数	60分	498*
短答式・語彙読解	短答式	61	正誤	正答数	40分	500
語彙	多肢選択	30	正誤	正答数	15分	500

*無解答が多かった2名の受検者を除外した。

- 研究に参加した大学生は500人で，5つの客観テストをすべて受検した。このうち，無作為に選ばれた167人が論述テストも受検した。

論述テストの概要

- 論述テストは以下の3問で構成されている。

資料活用型

複数の資料（非連続テキスト主体）を根拠にし，定められた主張を論理的に展開する問題。

課題解決型

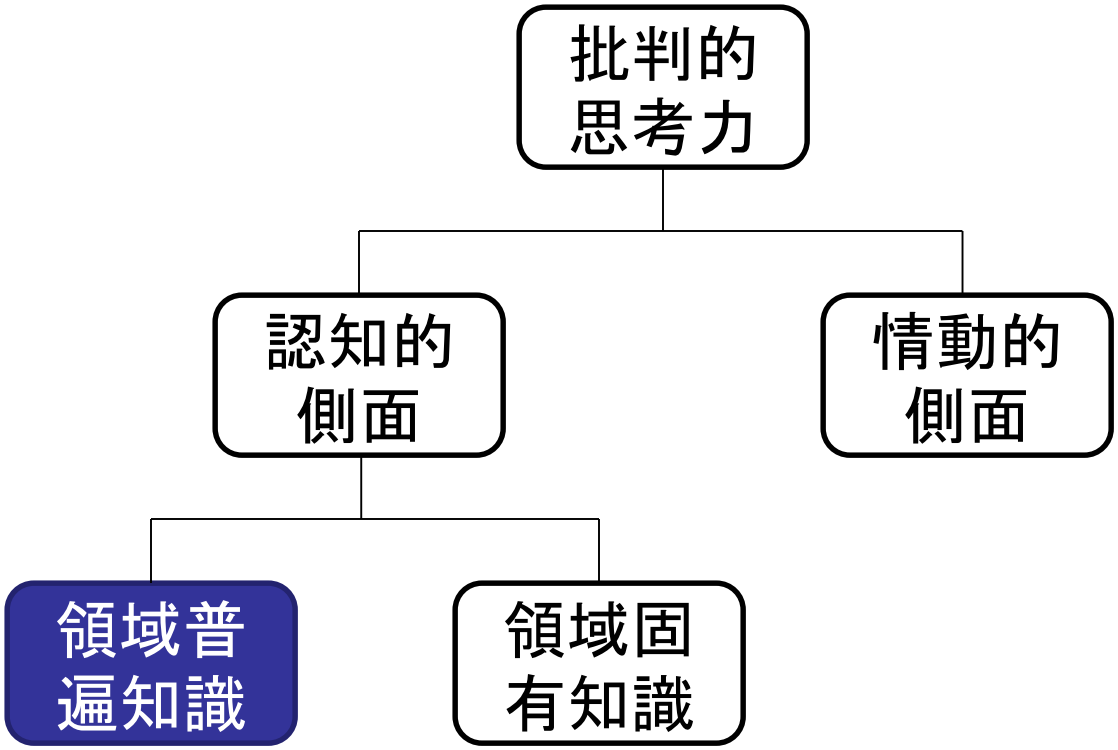
ある主張の根拠となっている複数の資料（非連続テキスト主体）を検討し，根拠として妥当でないことを論理的に指摘する問題。

反論型

ある特定の価値観を主張した資料（連続テキストのみで構成）の論構造を明確にし，論理的に反論する問題。

- どの問題においても書くべき要素が決まっており，それらの要素をどれだけ書けたかで5段階に評価された。

批判的思考テストの概要



- 批判的思考テストでは，「領域普遍知識」を構成する「情報の明確化」「情報の分析」「推論」（楠見・子安・道田，2011）の3スキルを測定している。

研究の枠組み

- 論述テストが論述力を正しく測定していると仮定する。論述テストの観測得点を古典的テスト理論に基づいて以下のように書く。

$$X_{論} = T_{論} + E_{論} \quad (1)$$

- この枠組みに従えば，論述テストの問題数を増やして測定誤差の分散を小さくすることが，論述力を正確に測定するための正しいアプローチになる。
  - しかし実際には，論述テストの問題数を増やすのは困難であるため，客観テストを併用することで，間接的に測定の正確さを上げることを考える。

研究の枠組み

- 客観テストpの観測得点を同様に古典的テスト理論に基づいて以下のように書く。

$$X_p = T_p + E_p \tag{2}$$

- 論述テストと客観テストの観測得点を足し合わせた複合得点（composite score）を以下のように書く。

$$X_{comp} = w_1 X_{論} + w_2 X_p \tag{3}$$

  - ここでw_1およびw_2はそれぞれ論述テスト，客観テストに対する重みである。特に明記しないが，観測得点は標準化されているものとする。

評価指標

- 古典的テスト理論の一般的な仮定に従うと，論述テストの真値と複合得点の相関係数は以下のように書ける。

$$\mathrm{cor}\left(T_{論}, X_{\mathrm{comp}}\right) = \frac{w_1\sqrt{\mathrm{rel}(X_{論})} + w_2\mathrm{cor}\left(T_{論}, T_p\right)\sqrt{\mathrm{rel}(X_p)}}{\sqrt{w_1^2 + w_2^2 + 2w_1w_2\mathrm{cor}\left(T_{論}, T_p\right)\sqrt{\mathrm{rel}(X_{論})}\sqrt{\mathrm{rel}(X_p)}}} \quad (4)$$

  - ここでcorは相関係数，relは信頼性係数を表すものをする。

- 信頼性係数が真値と観測得点の相関係数の2乗であることの類推として，(4)式を2乗したものを測定の正確さの評価指標として使用する。

各テストの信頼性と論述テストとの相関

テスト	一般化可能性係数（元の問題数）	論述テストとの相関係数	
		希薄化の修正前	希薄化の修正後
論述	.610 (3)	-	-
批判的思考	.780 (40)	**.593**	**.860**
汎用読解	.755 (25)	.531	.782
教科型読解	.691 (17)	.445	.685
短答式・語彙読解	.828 (61)	.459	.646
語彙	**.844** (30)	.440	.613

- 問題数を変えたときの信頼性を求めやすいように，一般化可能性理論を使用している。

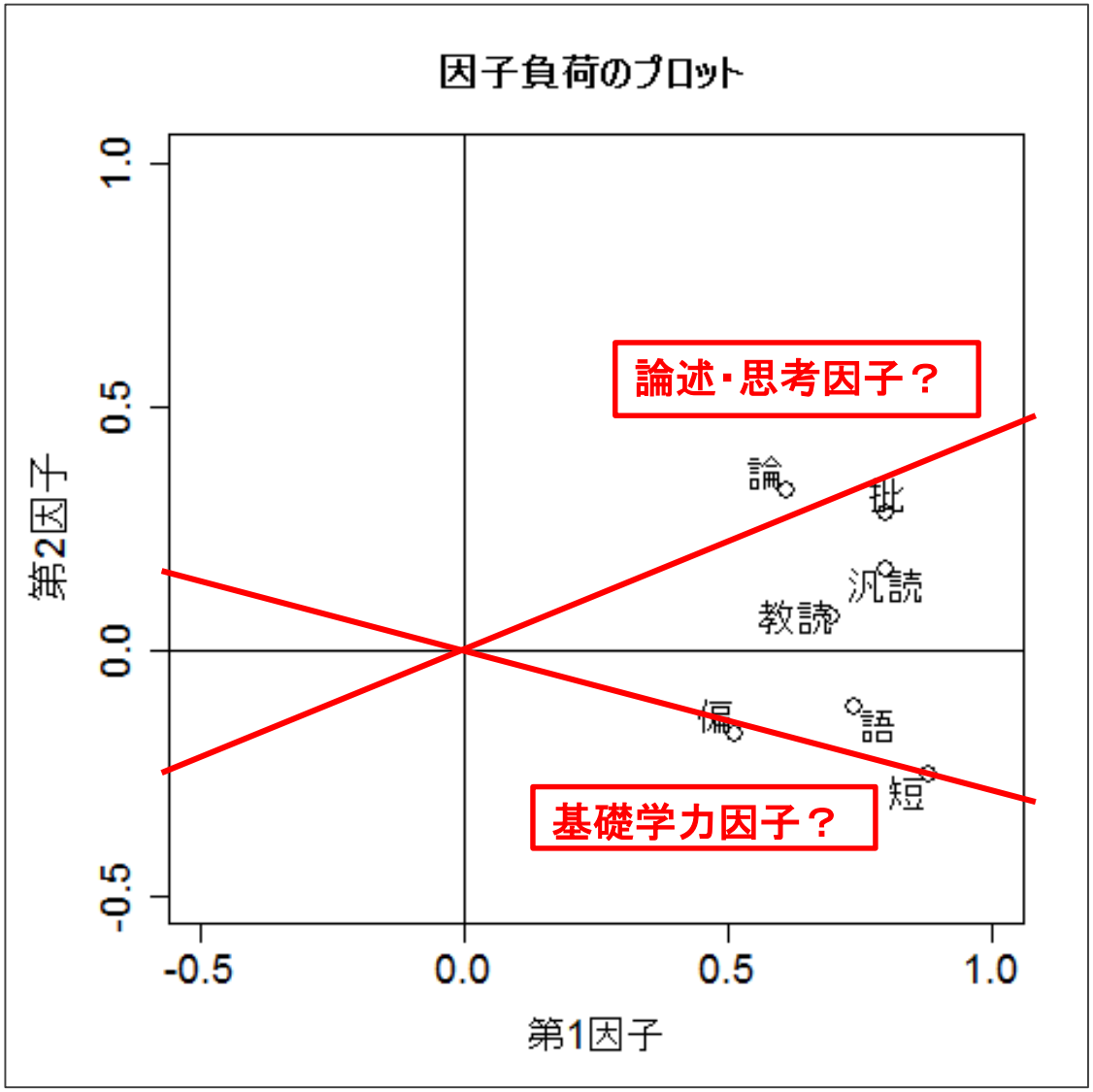
因子分析

- 6つのテストデータに，受検者が所属する大学の入学偏差値を変数として加え，7変数で因子分析を行った。

- Rのpsychパッケージ（Revelle, 2016）内のfa.parallel関数を使用した平行分析では，2因子モデルが妥当であると判断された。

- 同パッケージ内のfa関数を使用して，最尤法による因子分析を行った。

因子分析の結果



- 左図は回転前の因子負荷をプロットしたもの。
- 第1因子は各変数のeffective weight，第2因子は認知処理の複雑さを表しているように見える。
- オブリミン回転後の因子軸は赤線のようになる。

評価指標を使った分析

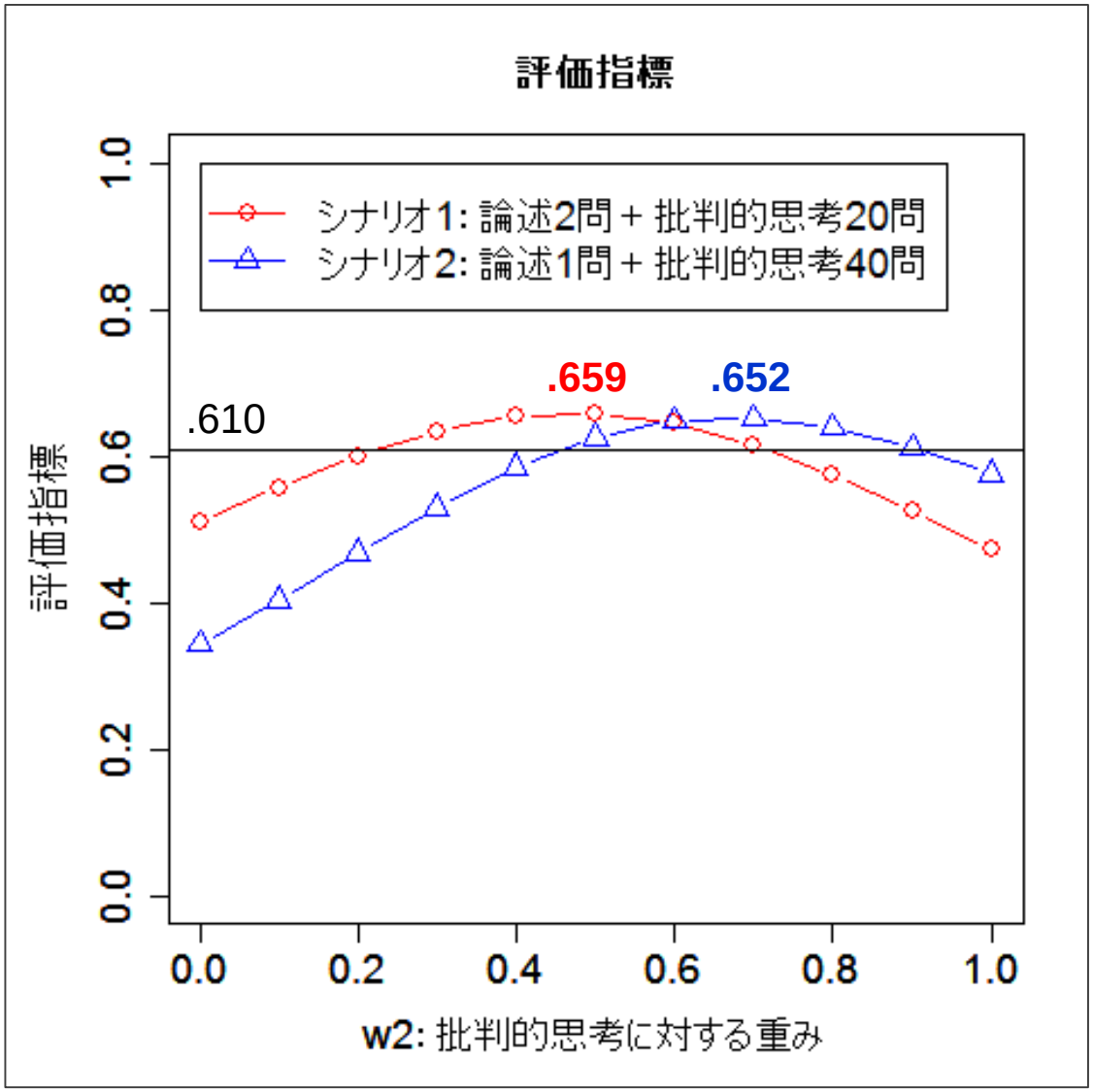
- ここまでに得られた結果から，批判的思考テストが最も論述力に近い能力を測定していることがわかった。

- そのため，評価指標を使った分析では，論述テストと批判的思考テストを組み合わせたときに，論述力の測定の正確さがどのように変化するかを調べる。

- また，現実的なテスト時間は90分程度であることから，以下の2つのシナリオについて比較を行った。

  シナリオ1：論述2問＋批判的思考20問
  シナリオ2：論述1問＋批判的思考40問
  ベースライン：論述3問

評価指標を使った分析



- $w_1 + w_2 = 1$とし，w_2を0から1まで0.1ずつ変化させたときの評価指標の変化を示したもの。
- 黒線は論述3問のときの評価指標の値を示している。
- シナリオ1は$w_2 = .5$，シナリオ2は$w_2 = .7$のときに評価指標が最大になった。

まとめ

- シナリオ１とシナリオ２を比較したところ，重みを適切に決めれば，どちらも論述力を同じ程度正確に測定できることが示唆された。
  - シナリオ2は論述問題が1問なので，採点のコストを考えると，シナリオ2のほうが好ましいことになる。
- しかし，シナリオ２で評価指標が最大になった時の値は.65程度であり，正確な測定ができているとは言いがたい結果だった。
- このことから，90分のテスト時間で論述力を測定することはかなり難しいことがわかる。

考察

- 今後検討すべき課題として，以下のことが挙げられる。
  - 論述力ではなくて論述・思考因子を測定したら？
  - 複数の客観テストを混合したら？
  - 段階的に論述を書かせ，そのステップを得点に含めたら？
- 今回の結果は，あくまで開発中のテストで見られた結果であり，広く妥当性が確立されたテストで得られた結果ではない。
- 論述力テストおよび各客観テストの妥当性の検証を進めるとともに，今回の結果が再現性のあるものなのかを並行して確認していく必要がある。

引用文献

楠見　孝・子安増生・道田泰司（編）（2011）．　批判的思考力を育む―学士力と社会人基礎力の基盤形成　有斐閣

Revelle, W. (2016). psych: Procedures for personality and psychological research [Computer software]. http://CRAN.R-project.org/package=psych. (R package version 1.6.6)

芝 祐順（1978）．　語彙理解尺度作成の試み　東京大学教育学部紀要　第17巻　pp.47-58.

発表は以上です。
ご清聴いただきありがとうございました。

問題例：情報の分析

早起きをする習慣がある人は寿命が長い、という研究結果が発表された。

ユニバーサル大学のコーレン教授は、長寿者が多いことで有名な南太平洋の孤島Zに住む男女約100名を対象として10年以上にわたる調査を行った。調査地は、長寿の秘密を解明する手がかりを見つけることを期待して選ばれたものだ。調査の結果から、ふだん午前5時よりも前に起きる習慣がある人は、そうでない人に比べて平均5.8年寿命が長いことがわかった。その他に、長寿者とそうでない人々の間で目立った他の生活習慣の違いはみられなかったという。コーレン教授は「早起きにこれほどの効果があるとは驚きだ」と述べている。

問：この研究成果に基づいて「早起きをする人は寿命が長い」という結論を導くことが適切かどうか判断するため、確認するべきことは何か。最もふさわしいものを一つ選びなさい。

① コーレン教授が所属する大学は、研究に力を入れている大学か。
② コーレン教授は、企業や省庁から研究のための予算提供を受けていなかったか。
③ 調査を行った10年のあいだに、Z島にどのくらいの人々の出入りがあったか。
④ 長寿ではない地域でも同じような結果が得られるか。
⑤ 調査の対象となった人々の就いている職業による寿命への影響は調べてあったか。

正答：４