【原　著】

400字論述課題における能力測定の信頼性

大久保智哉*

要　約

本研究において，400字論述課題の能力測定の信頼性について検討をおこなった．400字以内で論述させるような課題を受験者に課した場合，評価得点はどのような要因から構成され，どの程度の信頼性を持つものなのかという点が検討された．具体的には，受験者，評価者，課題といった3つの要因とそれらの交互作用を含めた要因の分散成分を推定することにより，能力測定の信頼性を明らかにした．また，調査から推定された分散成分からテストデザインに応じた信頼性係数を推定し，テストデザインごとの信頼性について報告した．

キーワード：論述課題，パフォーマンス評価，信頼性，一般化可能性理論

1　はじめに

大学入学者選抜において小論文試験がしばしば用いられている．小論文試験とは，受験者が与えられた課題について論述するという形式を持つものであるが，課題の内容やその採点基準，またはその採点方法は大きく試験によって異なっている．小論文試験では出題されたテーマについて制限時間内に受験者が論述したものを，採点者が決められた基準にしたがって採点することによって得点化するという形式となっている．一般的に，このようなテスト形式の評価はパフォーマンス評価と呼ばれている．現在，わが国の入学者選抜の場面において，小論文試験や面接試験といったパフォーマンス評価は多枝選択形式に次いで多く用いられているテスト形式であろう．パフォーマンス評価がその得点化の難しさや実施上のコストにも関わらず多く用いられているのは，多枝選択項目では測定できないと考えられている能力特性を測定していると考えられているからである．

米国においては，SAT（Scholastic Assessment Test；大学進学適性試験）における Essay Test や USMLE（United State Medical Licensing Examination；米国医師国家試験）における OSCE（Objective Structured Clinical Examination；客観的臨床能力試験）などの実技試験がパフォーマンス評価としてハイステークスな試験において採用されている．わが国においても大学入試において，アドミッション・ポリシーに基づく選抜方法の多様化，評価尺度の多元化が求められており（「大学入試の改善について」平成12年11月22日大学審議会），教科型の多枝選択型とは異なったテストの必要性が高まっている．

このような流れからも，大学入試のみならずハイステークスな試験において，現状よりもパフォーマンス評価の利用は増えていくと考えられる．しかしながら，わが国ではパフォーマンス評価をテスト形式として用いたときの知見が十分に蓄えられているとは言えない．本研究では，その中で入学者選抜の場面における利用を考えたときに有力な論述形式の課題を取り上げる．

わが国では，小論文試験においては一つの課題のみから構成されている．入学者選抜において，論述形式の項目が1つのみしか出題されない理由は，受験時と採点時の時間的制約にあると考えら

* 独立行政法人大学入試センター研究開発部
2013年3月1日　受理

れる．多くの大学では，小論文試験の試験時間を60 分に設定し，制限文字数を 800 字～1200 字として回答させている．なお，本研究において用いる論述型項目は従来の小論文試験よりも少ない文字数で論述させるため，400 字論述課題と呼ぶことにする．本研究において 400 字論述課題を採用した理由は，論述文字数を減らすことで，一つの回答作成に要する時間を減らし，その分で受験者一人あたりから複数の課題に対する論述回答を得ようとするためである．このような 400 字論述課題の妥当性については，大学入試センター試験で用いられる科目を外的基準として用いて検討した研究がある（荒井・石岡・宮埜, 印刷中）．

400 字論述課題において得点を構成する主な要因は受験者（people）と課題（task），評価者（rater）である．多枝選択形式のテストとは異なり，評価者が得点を構成する要因に加わる．一般的には，得点の構成要因が増えることが受験者に対する評価の信頼性を低める原因となる．そのため，実際には一つの評価対象に対して評価者を複数用意し，その平均値をその評価対象の得点として採用することが多い．これは，評価者を変動要因としてとらえた場合に，その要因内の水準の差が与える影響を抑えるための工夫である．ここで，適切な評価者の人数について検討される必要がある．評価者が多い方が評価者要因の得点に対する影響を抑えることができるものの，評価者を増やすことはコストを増やすことになるため，その数はできるだけ抑えることがのぞまれている．人数による信頼性の変化に関する知見があれば，テスト実施時に適切な評価者数の判断に役立つであろう．なお，論述形式項目の信頼性において，評価者間の相関係数を採用する場合があるがこれは不適切である．これは，受験者と課題の分散成分が真値と交絡し，信頼性の過大推定につながるためである (Brennan, 2000)．実際，評価者間の相関係数については高い傾向を示した報告が多く，また評価者の分散成分についても比較的小さいことが確認されているものの，交互作用項の分散成分については有意に大きいことが示されており，信頼性を過大評価していることが指摘されている (Brennan, 1996)．具体的には，このことは課題によって受験者の順位が入れ替わってしまうことがあるということを意味している．

また，入学者選抜におけるパフォーマンス評価では，評価者が全ての回答を評価することは不可能である．したがって，評価者は全体の回答の一部（評価ブロックと呼ぶ）を評価することになる．パフォーマンス評価においては，評価者には評価に際して，評価対象内での相対的な評価をさせることが多い．しかし，そのような評価方法をとると，評価ブロック間での得点の等質性が失われてしまう．すなわち，評価ブロック間での得点が比較可能でなくなってしまうのである．具体的には，評価ブロック間で評価対象の真の得点に偏りがある場合には，得られた得点を同一尺度上で表すことが不適切となる．

このように，パフォーマンス評価には，多枝選択式項目とは異なり，評価に際して十分な検討が必要となっている．そこで，本研究においては

1. 試験時間内に複数の課題を与えることができる 400 字論述課題を提案し，日本の大学 1 年生に対して，複数の課題を与え，複数の評価者によって評価するという完全クロス計画によって評価データを収集した上で，その能力測定の信頼性について検討する．
2. 評価者に対して，あらかじめ評価対象の得点の分布を指定しない方法によって，採点をさせる．そのことによって，能力測定の信頼性にどのような影響を与えるかを検討する

ということを目的とする．

2 400 字論述課題の実施

本研究の 400 字論述課題は平成 24 年度大学入試センターモニター試験内においておこなわれた．実験日は平成 24 年 1 月 21 日であった．

2.1 実験デザイン

本研究では，受験者群を 2 群に別け，それぞれの群に対して 2 つの論述形式の課題を課した．第 1 群に対しては，課題 1 と課題 2 を，第 2 群に対しては，課題 1 と課題 2′ を回答してもらった．したがって，受験者は 2 問に回答したことになる．

試験時間は 60 分であったが，それぞれの課題に対して回答時間は 30 分と設定した．途中の休憩はなく，前半の 30 分で課題 1 を回答してもらい，後半の 30 分で課題 2，もしくは課題 2′ を回答して

もらった．前半の回答時間に後半の課題を回答する，もしくは後半の回答時間に前半の課題を回答することがないように教示を与え，試験時間中には試験監督が見回った．なお，本研究では課題 2 と課題 2′ を区別しない．課題 2 と課題 2′ の等質性については荒井・石岡・宮埜（投稿中）を参照にされたい．

回答用紙には，20 字 × 20 行の横書きの回答用紙を用いた．

2.2 受 験 者

受験者は，平成 24 年度大学入試センターモニター試験に参加した 213 名の大学 1 年生であった．受験者の所属する大学は都内近郊である．第 1 群の人数は 108 名で，第 2 群の人数は 105 名で，合計 213 名であった．第一群と第二群は無作為に割り当てた．

2.3 課　　題

本研究で用いた課題 1，2，2′ は次のようなものであった．

課題 1　同じ内容の発明が複数ある場合に，そのいずれに対して特許を付与するかを決定する原則として以下の二つがある．

- 先願主義：先に出願した者に特許を付与する主義
- 先発明主義：先に発明した者に特許を付与する主義

この二つの主義について，あなたはどちらが特許を与える原則としてふさわしいと考えますか．ふさわしいと思う主義を示した上で，その考えに至った理由をあげて，**400 字以内**で記述しなさい．

課題 2　言葉は，しばしば，辞書等で本来の意味とされる意味（正用）のほかに，本来とは異なるとされる意味（誤用・慣用）でも使われます．次の表は，そのような言葉の例です．

言葉	本来の意味	本来とは異なる意味
情けは人のためならず	人に情けを掛けておくと，巡り巡って結局は自分のためになる	人に情けを掛けて助けてやることは，結局はその人のためにならない
雨模様	雨が降りそうな様子	小雨が降ったりやんだりしている様子

このような言葉の誤用・慣用について，あなたの考えを **400 字以内**で論じなさい．

課題 2′　言葉は，しばしば，辞書等で本来の意味とされる意味（正用）のほかに，本来とは異なるとされる意味（誤用・慣用）でも使われます．次の表は，そのような言葉の例です．

言葉	本来の意味	本来とは異なる意味
情けは人のためならず	人に情けを掛けておくと，巡り巡って結局は自分のためになる	人に情けを掛けて助けてやることは，結局はその人のためにならない
雨模様	雨が降りそうな様子	小雨が降ったりやんだりしている様子

このような言葉の誤用・慣用を，自分自身が使うかどうかについては，

- どのような場合でも使おうとは思わない
- 相手や状況によっては使うと思う

など，様々な考えがあります．あなたは，言葉を本来とは異なる意味（誤用・慣用）で使いますか．あなたの考えを示した上で，その理由をわかりやすく **400 字以内**で説明しなさい．

アンケート　さらに，本研究では，小論文の経験の有無について下記のように回答させた．

論文あるいは小論文の経験について，当てはまるものに○をつけて下さい．

1. 大学受験の際の受験科目に論文あるいは小論文が含まれていた．
   はい・いいえ
2. 大学の課題として論文あるいは小論文がある．
   はい・いいえ

3 400 字論述課題における評価方法

評価者による採点にあたって，下記の内容をインストラクションとして与えた．なお，インストラクションは下記内容が記述されたものを配布することによっておこなった．

3.1 採点に際しての指示

採点は次の手順に沿っておこなう．下記の手順から逸脱したものは採点として認められない．複数課題を採点する場合でも，必ず課題ごとに採点

指示書に従うこと．

1. 採点指示書の理解
   (a) 採点指示書を理解するまで読む．
   (b) 不明な点がある場合には指示者へ質問すること．それでも理解できない場合は採点を辞退すること．
2. 採点について
   (a) 採点は 4 つの観点，「課題に対する論点の一致性」，「論理の明瞭さ」，「表現や言葉の適切な使用」，「総合評価」についておこなう．
   (b) 回答の水準によって 5 段階での採点をおこなう．0 点から 4 点である．
   (c) 基準に従って採点をおこなう．相対評価ではなく，絶対評価である．
   (d) 採点観点は「指定された場所にできるだけ適切な材料によってまっすぐな塔を建てられるか」に等しいと理解して欲しい．「指定された場所に近いか」は「課題に対する論点の一致性」において採点をする．また，「適切な材料を用いているか」は「表現や言葉の適切な使用」において評価する．また，「塔がまっすぐか」については「論理の明瞭さ」において採点をする．また，完成した塔そのものについては「総合評価」において採点する．
3. 回答の前読み
   (a) ある課題について採点を始める前に，全ての受験者の回答を採点の前に読む．何回読んでもかまわない．
   (b) 前読みの目的は，採点者が採点する際の基準を事前に確認するためである．
   (c) 前読みの段階で疑問がある場合には，指示者へ質問すること．
4. 採点
   (a) 用意されたエクセルファイルに採点を記入していく．採点は与えられた回答順でおこなう．
5. 採点内容の確認
   (a) 自ら採点した内容について最低，一回は確認する．すなわち，一つの回答に対して，前読み，採点，確認と 3 回は読むことになる．

3.2 採点の基準

1.「課題に対する論点の一致性」：与えられた課題と回答者の論点が一致しているかという点について評価する．与えられた課題に対して，論点がずれている場合には評価が低くなる．
   0 点：課題と回答の論点がほぼ合っていない．
   1 点：課題と回答の論点が合っていない部分が多いが，論点が合っている部分が若干ある．
   2 点：課題と回答の論点が合っていない部分も合っている部分もおおよそ半々である．
   3 点：課題と回答の論点が合っている部分が多いが，論点が合っていない点が若干ある．
   4 点：課題と回答の論点がほぼ合っている．
2.「論理の明瞭さ」：回答者の論述内容が評価者にわかりやすく文章となっているかを評価する．論点が何度も変わる場合や論理に矛盾がある場合には評価が低くなる．論述内容が単純か複雑かを評価するのではない．
   0 点：論述全体においてほぼ論理が明瞭でない．
   1 点：論理が明瞭でない部分が多いが，論理が明瞭である部分が若干ある．
   2 点：論理が明瞭でない部分と明瞭である部分がおおよそ半々である．
   3 点：論理が明瞭である部分が多いが，論理が明瞭でない部分が若干ある．
   4 点：論述全体においてほぼ論理が明瞭である．
3.「表現や言葉の適切な使用」：論述に用いている表現や言葉が適切に用いられているかについて評価する．論述の構成を評価するのではない．
   0 点：表現や言葉が適切に用いられていない．
   1 点：表現や言葉が適切に用いられていない部分が多いが，適切に用いられている部分が若干ある．
   2 点：表現や言葉が適切に用いられていない部分と用いられている部分がおおよそ半々である．
   3 点：表現や言葉が適切に用いられている部分が多いが，適切に用いられていない部分が若干ある．
   4 点：表現や言葉が適切に用いられている．
4.「総合評価」：400 字の論述について全体を評価する．全体としてどのような評価を与えるかという観点から採点する．

0 点：論述文の水準として低い．
1 点：論述文の水準としてやや低い．
2 点：論述文の水準としては高くも低くもない．
3 点：論述文の水準としてやや高い．
4 点：論述文の水準として高い．

3.3 評 価 者

評価をおこなった 3 名は，大学院博士課程在学中の学生 2 名と修士課程在学中の学生 1 名であった．なお，修士課程在学中の学生については，学術論文執筆経験がある者であった．

4 評価の結果

4.1 採点結果

本研究によって得られた採点結果についてそれぞれ示す．

本研究では，採点指示の段階で言わば絶対評価のような基準を設けて採点をさせたため，得点分布については大きく低得点側へ歪んだものもある．たとえば，慣用表現についての課題の論点の一致性を評価させたものは，評価者 1〜3 いずれにおいても大きく得点分布が低得点側へと歪んでいる．これは，課題 2 において考えを論じるように課題として指示したが，ほとんどの学生が「慣用表現を用いることに賛成か反対か」を論じてしまったため，評価者によって論点の一致がないと判断されてしまったためである．このような状況においては，相対的な評価をおこなった方が評価得点を利用しやすいものとなるが，一方でその値は評価ブロック間での等質性が保証されないことになる．

4.2 分散成分の推定結果

本研究で得られた評価結果をもとに一般化可能性理論 (Shavelson & Webb, 1991; Ruiz-Primo, et al., 1993; 池田央, 1994; Cronbach, et al., 1997; Brennan, 2001) の枠組みで信頼性の評価をおこなった．まず，本研究において得られた 400 字

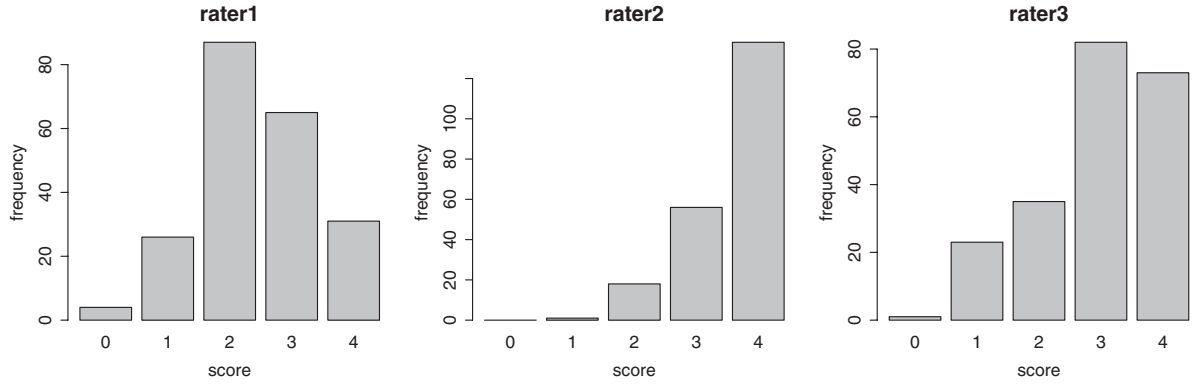

図 1 ［特許について］［課題に対する論点の一致性］についての得点分布

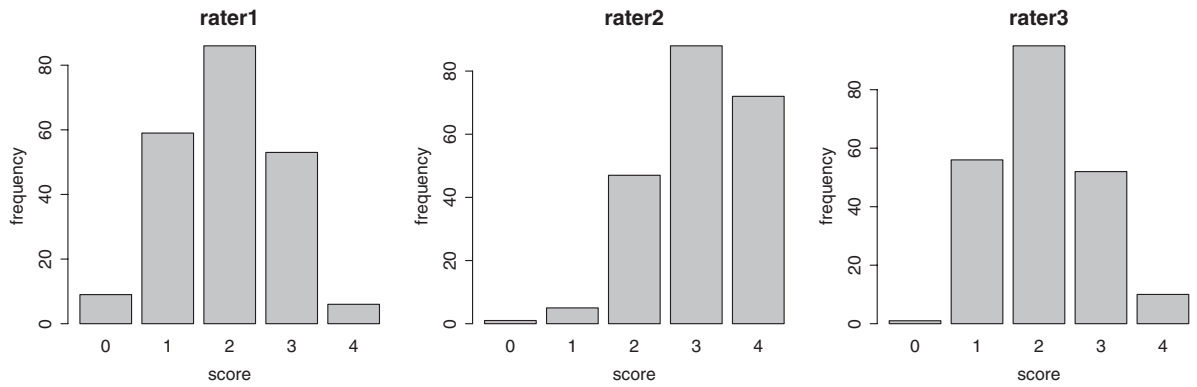

図 2 ［特許について］［論理の明瞭さ］についての得点分布

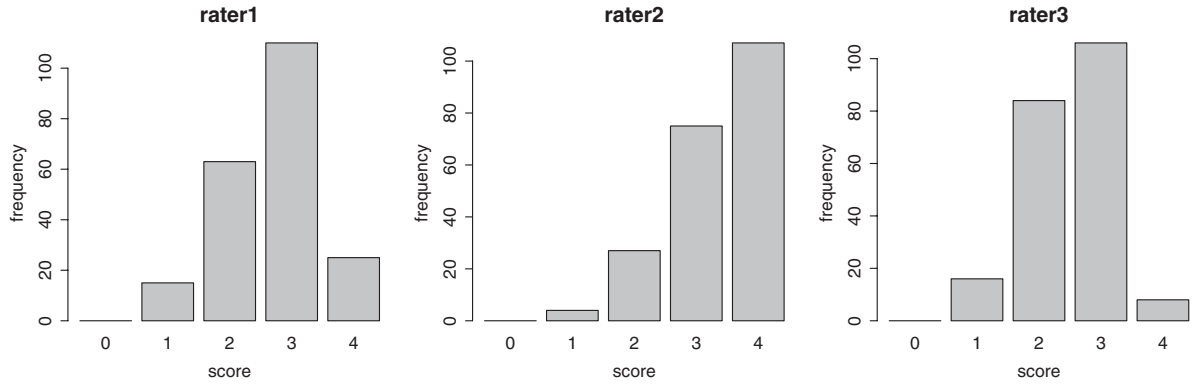

図 3 ［特許について］［言葉や表現の適切さ］についての得点分布

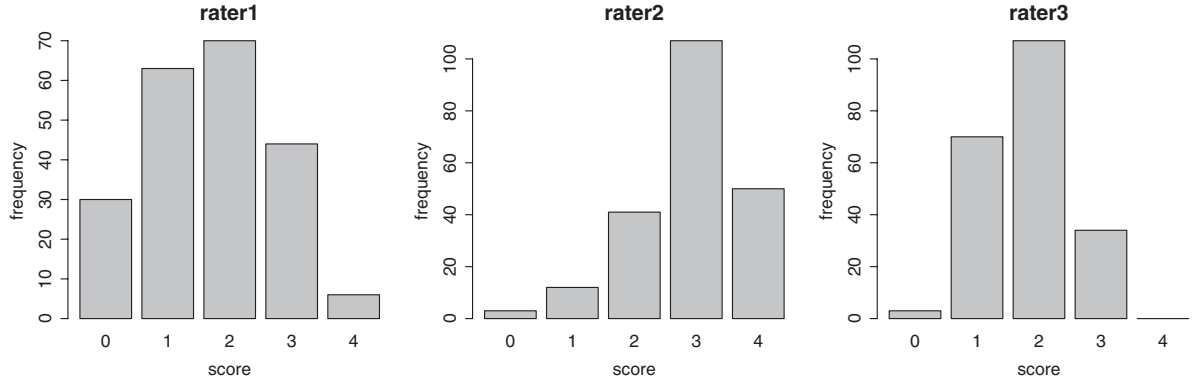

図 4 ［特許について］［総合評価］についての得点分布

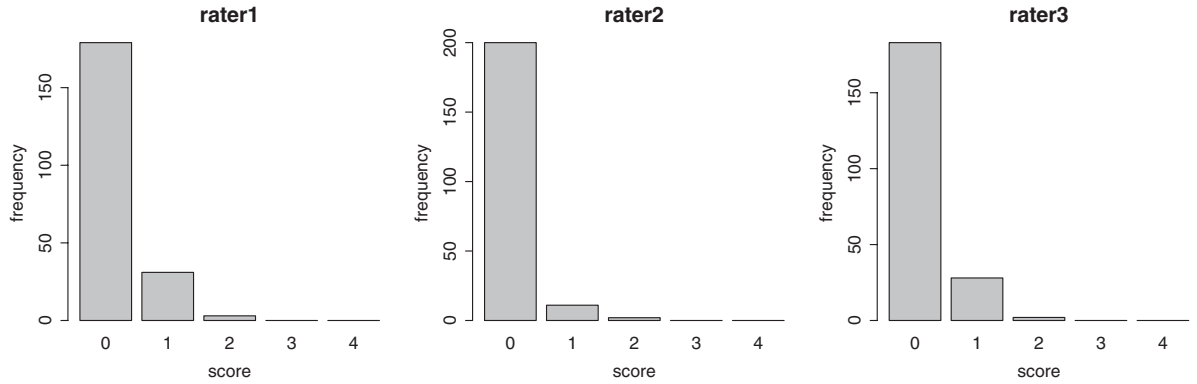

図 5 ［慣用表現について］［課題に対する論点の一致性］についての得点分布

論述課題の答案を採点したデータを混合モデル (Fitzmaurice, Laird, & Ware, 2011) を用いて各要因の分散成分の推定をおこなった.

具体的にモデルに入れた要因は下記の 7 つである. また, これらはすべて変量効果としてモデルに組み入れられた.

1. 受験者（people）
2. 評価者（rater）
3. 課題（task）
4. 受験者 × 評価者（people × rater）
5. 受験者 × 課題（people × task）
6. 評価者 × 課題（rater × task）
7. 誤差（error）

なお, 推定に際しては制限付き最尤推定法(Restricted Maximum Likelihood Estimation, REML) を用いた.

また, 本研究において分散成分の推定に用いたデータは以下のものである.

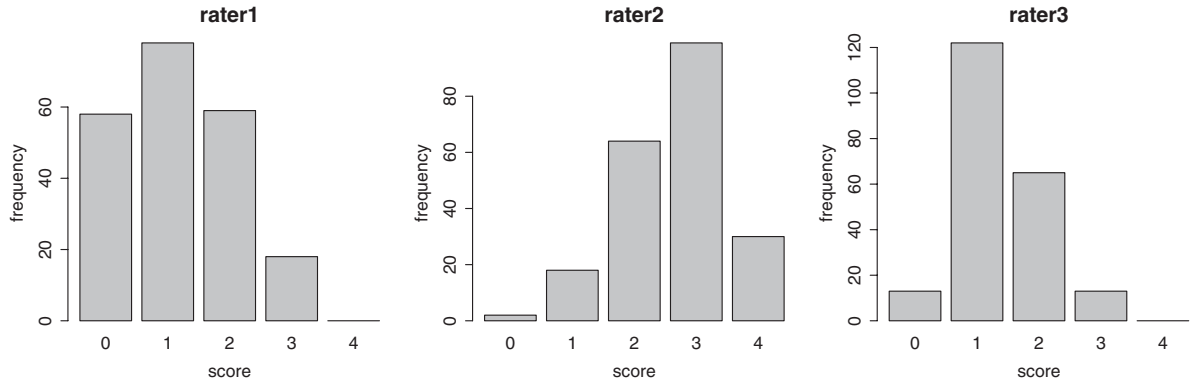

図 6 ［慣用表現について］［論理の明瞭さ］についての得点分布

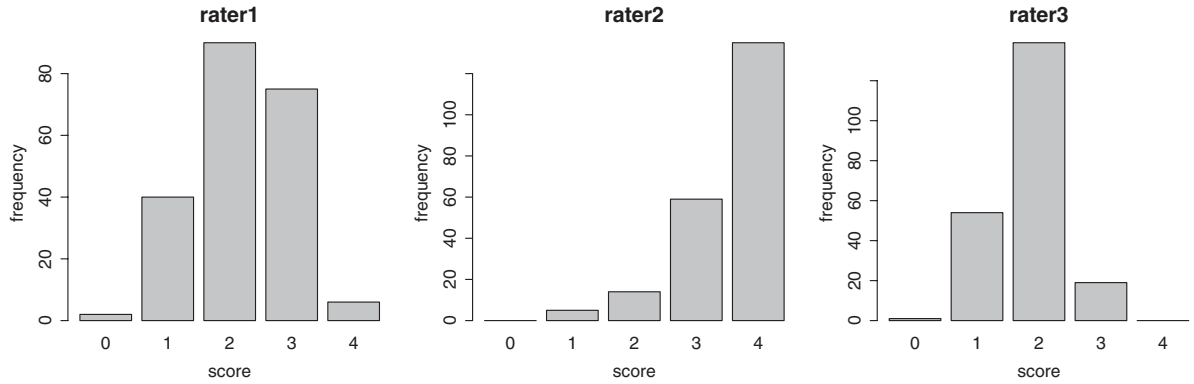

図 7 ［慣用表現について］［言葉や表現の適切さ］についての得点分布

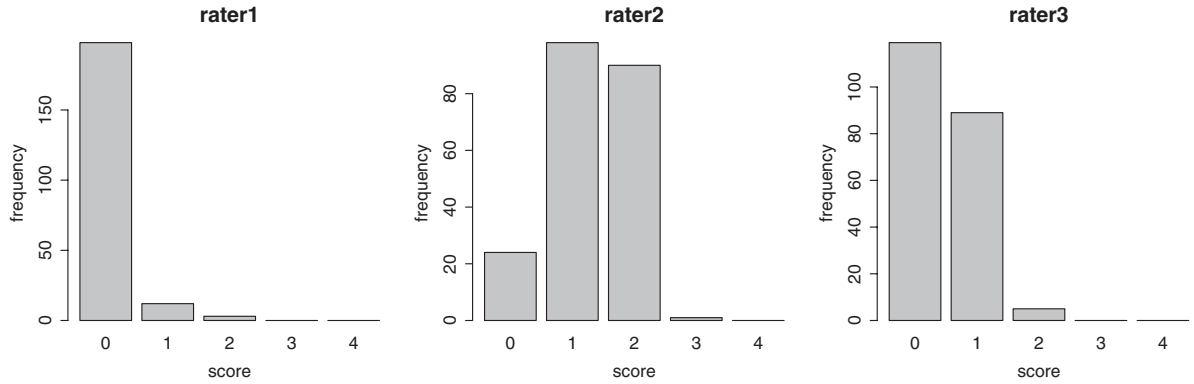

図 8 ［慣用表現について］［総合評価］についての得点分布

1.「課題に対する論点の一致性」についての採点データ
2.「論理の明瞭さ」についての採点データ
3.「表現や言葉の適切な使用」についての採点データ
4.「総合評価」についての採点データ

それぞれについて推定された分散成分を示す．「課題に対する論点の一致性」についてのモデルは不適解が得られた．評価者に関わる項において分散成分の推定値が 0 となった．その他の観点については，推定値が得られた．

得られた推定値についてであるが，全体的に課題や評価者の分散成分が大きいことが確認された．一方で受験者の分散成分は大きくなく，受験者を識別するという観点からは望ましくない結果が得られた．これは採点指示書を変えることによって改善することができるかもしれないが，それは今後の研究が待たれよう．

表 1 「課題に対する論点の一致性」についての分散成分推定値

成分	分散比	分散成分	標準誤差	下側 95%信頼限界	上側 95%信頼限界
受験者	0.025	0.008	0.017	−0.024	0.041
評価者	0.000	0.000	0.000	0.000	0.000
課題	12.444	4.006	5.740	−7.245	15.256
評価者 × 受験者	0.000	0.000	0.000	0.000	0.000
課題 × 受験者	0.388	0.125	0.023	0.080	0.170
課題 × 評価者	0.486	0.156	0.112	−0.062	0.375
残差		0.322	0.016	0.291	0.353

表 2 「論理の明瞭さ」についての分散成分推定値

成分	分散比	分散成分	標準誤差	下側 95%信頼限界	上側 95%信頼限界
受験者	0.164	0.077	0.028	0.023	0.131
評価者	1.035	0.487	0.496	−0.486	1.460
課題	0.404	0.190	0.278	−0.355	0.736
評価者 × 受験者	0.097	0.046	0.025	−0.004	0.095
課題 × 受験者	0.258	0.122	0.029	0.065	0.179
課題 × 評価者	0.033	0.016	0.018	−0.019	0.051
残差		0.471	0.032	0.407	0.534

表 3 「表現や言葉の適切な使用」についての分散成分推定値

成分	分散比	分散成分	標準誤差	下側 95%信頼限界	上側 95%信頼限界
受験者	0.160	0.060	0.019	0.024	0.097
評価者	1.042	0.394	0.447	−0.482	1.270
課題	0.048	0.018	0.080	−0.139	0.176
評価者 × 受験者	0.154	0.058	0.021	0.016	0.100
課題 × 受験者	0.086	0.033	0.018	−0.002	0.067
課題 × 評価者	0.260	0.098	0.100	−0.098	0.295
残差		0.378	0.026	0.327	0.429

表 4 「総合評価」についての分散成分推定値

成分	分散比	分散成分	標準誤差	下側 95%信頼限界	上側 95%信頼限界
受験者	0.100	0.037	0.021	−0.004	0.078
評価者	1.122	0.409	0.414	−0.403	1.221
課題	3.087	1.125	1.597	−2.005	4.255
評価者 × 受験者	0.015	0.006	0.018	−0.030	0.041
課題 × 受験者	0.356	0.130	0.026	0.079	0.180
課題 × 評価者	0.023	0.009	0.010	−0.012	0.029
残差		0.365	0.025	0.315	0.414

4.3 テストデザインごとの信頼性係数の推定

次に，得られた分散成分の推定値をもとに信頼性係数を推定する．本研究で用いた信頼性係数は Generalizability 係数（ρ^2）であり，下記で表される．

$$\rho^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pt}^2}{n_t} + \frac{\hat{\sigma}_{pr}^2}{n_r} + \frac{\hat{\sigma}_e^2}{n_t n_r}} \qquad (1)$$

ここで，$\hat{\sigma}_p^2$ は受験者の分散成分推定値，また $\hat{\sigma}_{pt}^2$，$\hat{\sigma}_{pr}^2$，$\hat{\sigma}_e^2$ はそれぞれ受験者 × 課題，受験者 × 評価者，誤差の分散推定値を表している．また，n_r は評価者の数，n_t は課題の数を表す．

得られた分散成分の推定値をもとに，評価者と課題の数をそれぞれ 1 から 10 に変化させた場合の信頼性係数 ρ^2 を推定したものが，下記の表である．

この結果より，論述課題の能力測定の信頼性を

上げることの難しさが示唆された．「論理の明瞭さ」と「表現や言葉の適切な使用」においては，課題数と評価者数を増やせば比較的良好な信頼性係数が得られることがわかったが，「論点の一致性」と「総合評価」では評価者数や課題数を増やしてもなかなか信頼性係数が上がらないことが示された．「論点の一致性」については，課題 2 において多くの回答が 0 点を与えられたことに起因すると思われる．また，「総合評価」についてはその評価基準が曖昧であるためであると考えられる．このことは，評価者から評価開始前に質問として受けていた．

表 5 「課題に対する論点の一致性」に関する Generalizability 係数（ρ^2）

		評価者数									
		1	2	3	4	5	6	7	8	9	10
課題数	1	0.02	0.03	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05
	2	0.04	0.05	0.07	0.07	0.08	0.08	0.09	0.09	0.09	0.09
	3	0.05	0.08	0.10	0.11	0.11	0.12	0.13	0.13	0.13	0.13
	4	0.07	0.10	0.12	0.14	0.15	0.15	0.16	0.17	0.17	0.17
	5	0.08	0.12	0.15	0.17	0.18	0.19	0.19	0.20	0.20	0.21
	6	0.10	0.15	0.17	0.19	0.21	0.22	0.22	0.23	0.23	0.24
	7	0.11	0.17	0.20	0.22	0.23	0.24	0.25	0.26	0.26	0.27
	8	0.13	0.19	0.22	0.24	0.26	0.27	0.28	0.28	0.29	0.29
	9	0.14	0.20	0.24	0.26	0.28	0.29	0.30	0.31	0.31	0.32
	10	0.15	0.22	0.26	0.28	0.30	0.31	0.32	0.33	0.34	0.34

表 6 「論理の明瞭さ」に関する Generalizability 係数（ρ^2）

		評価者数									
		1	2	3	4	5	6	7	8	9	10
課題数	1	0.11	0.17	0.21	0.23	0.25	0.27	0.28	0.29	0.30	0.31
	2	0.18	0.28	0.33	0.37	0.40	0.42	0.43	0.45	0.46	0.46
	3	0.24	0.35	0.42	0.46	0.49	0.51	0.53	0.54	0.55	0.56
	4	0.28	0.41	0.48	0.52	0.55	0.57	0.59	0.60	0.61	0.62
	5	0.32	0.45	0.52	0.56	0.60	0.62	0.63	0.65	0.66	0.67
	6	0.35	0.48	0.56	0.60	0.63	0.65	0.67	0.68	0.69	0.70
	7	0.37	0.51	0.58	0.63	0.66	0.68	0.70	0.71	0.72	0.73
	8	0.39	0.53	0.61	0.65	0.68	0.70	0.72	0.73	0.74	0.75
	9	0.41	0.55	0.62	0.67	0.70	0.72	0.74	0.75	0.76	0.77
	10	0.42	0.57	0.64	0.69	0.71	0.74	0.75	0.76	0.77	0.78

表 7 「表現や言葉の適切な使用」に関する Generalizability 係数（ρ^2）

		評価者数									
		1	2	3	4	5	6	7	8	9	10
課題数	1	0.11	0.19	0.25	0.30	0.34	0.36	0.39	0.41	0.43	0.44
	2	0.19	0.30	0.38	0.44	0.48	0.51	0.54	0.56	0.58	0.60
	3	0.24	0.37	0.46	0.51	0.56	0.59	0.62	0.64	0.66	0.67
	4	0.27	0.42	0.51	0.57	0.61	0.64	0.67	0.69	0.71	0.72
	5	0.30	0.45	0.54	0.60	0.64	0.68	0.70	0.72	0.74	0.75
	6	0.32	0.48	0.57	0.63	0.67	0.70	0.73	0.75	0.76	0.77
	7	0.34	0.50	0.59	0.65	0.69	0.72	0.74	0.76	0.78	0.79
	8	0.36	0.52	0.61	0.66	0.71	0.74	0.76	0.78	0.79	0.81
	9	0.37	0.53	0.62	0.68	0.72	0.75	0.77	0.79	0.80	0.82
	10	0.38	0.54	0.63	0.69	0.73	0.76	0.78	0.80	0.81	0.82

表 8 「総合評価」に関する Generalizability 係数（ρ^2）

		評価者数									
		1	2	3	4	5	6	7	8	9	10
課題数	1	0.07	0.10	0.13	0.14	0.15	0.16	0.17	0.17	0.18	0.18
	2	0.13	0.19	0.22	0.25	0.26	0.28	0.29	0.29	0.30	0.30
	3	0.18	0.26	0.30	0.33	0.35	0.36	0.37	0.38	0.39	0.40
	4	0.22	0.31	0.36	0.39	0.41	0.43	0.44	0.45	0.46	0.47
	5	0.26	0.36	0.41	0.45	0.47	0.48	0.50	0.51	0.51	0.52
	6	0.29	0.40	0.46	0.49	0.51	0.53	0.54	0.55	0.56	0.56
	7	0.32	0.44	0.49	0.53	0.55	0.57	0.58	0.59	0.60	0.60
	8	0.35	0.47	0.52	0.56	0.58	0.60	0.61	0.62	0.63	0.63
	9	0.38	0.49	0.55	0.59	0.61	0.62	0.64	0.64	0.65	0.66
	10	0.40	0.52	0.58	0.61	0.63	0.65	0.66	0.67	0.68	0.68

5 考　察

本研究において，400 字論述課題の評価に対する能力測定の信頼性について検討された．また，テストのデザイン（評価者の数，項目の数）を変えた場合の信頼性についても推定された．

推定された分散成分をみると，交互作用項についても大きい値が推定されており，単純に評価者間相関係数をもって能力測定の信頼性を主張することの危うさが改めて示された．また，同様の結果から，項目数（課題数）は論述形式の場合は多く必要であるということも示唆された．これは，課題と課題との交互作用項の分散成分が大きいことから理解することができる．したがって，一つの論述形式の項目をおこなうよりも，複数の項目を課すことが信頼性を高くする際の条件であることが本研究において改めて示された．これまで，大学入学者選抜に用いられている方法は，800 字から 1200 字の論述課題を 1 問だけ課すというものであるが，本研究からは，信頼性の観点からはこのような方法ではなく，複数の論述課題を課す方が適切であるということがわかった．その観点からも，400 字論述課題は有用であると考えられる．

また，結果から採点する観点によって信頼性係数が大きく異なることが示唆された．たとえば，「論点が一致しているか」という採点基準では多くの受験者が正しく「慣用表現」に関する課題文を理解できなかったために低得点を採ることになった．また，そのことが受験者要因の分散成分を小さくし，一方で課題間の分散成分を大きくする原因となった．このことは，課題選択の難しさ，リード文の重要性を示すものである．また，絶対的な評価によって採点をおこなったために，大きく得点分布が低得点側へ歪むことになり，そのことが受験者以外の要因の分散成分を大きくしてしまう結果となった．これらは，絶対的な評価を求めることの難しさであり，負の側面であるが，一方で相対的な評価を指示する場合には，採点時の標本内でのみ比較可能な評価となってしまうため，採点者が多くの受験者の採点をおこなうことができないような場合には採用することができない．また，年度間比較なども極めて難しくなってしまう．したがって，作題の観点から考察すれば，相対的な評価を評価者に課すか，絶対的な評価を課すかはテストの目的によるものの，絶対的な評価を課す場合には，課題の選択と誤解のないリード文の作成が非常に重要になってくるということが言えるであろう．しかし，採点指示書や課題自体を変えることによって改善できる部分もあるであろうが，そもそもこれらの結果は論述形式課題による受験者の特性値の推定の難しさを示したものであるとも言える．

さらに，「総合評価」という評価観点についても今後検討が必要であると思われる．すなわち，本研究では「総合評価」が評価者によって意味する内容が異なっていたと思われるからである．また，ある評価者からは「字の美醜」が影響を与えた可能性が指摘された．さらに，先行研究においても，渡部・曹 (1992) においてもその影響が論じられている．したがって，この「総合評価」については事前にその測定内容を評価者間で共有しておく必

要があるであろう．

これらの結果より，いずれの観点においても 0.7 の ρ^2 を得るには，多くのコストをかけないといけないことが示唆された．論点の一致性については，不適解が得られてしまったため，ここでは解釈をおこなわないが，他のいずれの観点においても，400 字を 30 分を書かすような論述課題を複数項目用意し，さらにそれらを複数（5 名以上）の評価者が評定をおこなわないと，信頼のできる測定とはならないことが示唆されている．本研究において，論述型項目をテストで用いることの難しさが示されたと言える．

本研究では，課題を 2 つに絞り，また評価者を 3 人の状況で分散成分を推定している．課題については今後増やすことはできないが，評価者については増やすことも可能である．したがって，今後の研究として，評価者を増やして再度分散成分の推定をおこなうことが必要であると思われる．さらに，採点指示書の見直しについても検討し，小論文試験等の論述形式の項目を用いる場合の作題上の知見を集めていく必要があると考える．

参考文献

荒井清佳・石岡恒憲・宮埜壽夫（印刷中）. 大学入学者選抜における小論文試験と教科・科目試験との関連について. 日本テスト学会誌.

池田央 (1994). 現代テスト理論（行動計量学シリーズ）. 朝倉書店.

渡部洋・曹亦薇 (1992). 小論文における字の美しさの影響について. 東京大学教育学部紀要. **32**, pp.253–256.

Brennan, R. L. (1996). Generalizability of performance assessments. In G. W. Phillips (Ed.). *Technical issues in performance assessments.*

Brennan, R. L. (2000). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice*, **19**(1), pp.5–20.

Brennan, R. L. (2001). *Generalizability Theory.* Springer.

Cronbach, L., et al. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, **57**(3), pp.373–399.

Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. (2011). *Applied Longitudinal Analysis (Wiley Series in Probability and Statistics)* 2nd Edition. Wiley.

Ruiz-Primo, M. A., et al. (1993). On the stability of performance assessments. *Journal of Educational Measurement*, **30**(1), pp.41–53.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer (Measurement Methods for the Social Science)*. Sage Publications, Inc.

Measurement reliability of 400-word essay

OKUBO Tomoya*

Abstract

This paper investigates measurement reliability of 400-word essay. Scores of the 400-word essay consist of three factors and their interactions: Ability of respondents, traits of raters, and characteristics of tasks. Variance components of the factors are estimated in this paper in order to research on reliability of the 400-word essay. Further, generalizability coefficients are estimated based on the estimated variance components, which enables us to predict the reliability of the 400-word essay according to test-designs.

* The National Center for University Entrance Examinations