課題レポート評価における評定者間の一致

阿久津洋巳*・嶋野恵美子**・熊谷　賢**・佐々木和歌**

(2005年2月7日受領)

Hiromi AKUTSU, Emiko SHIMANO, Satoshi KUMAGAI, Waka SASAKI

Agreement between Raters in Evaluating Homework Essays

はじめに

筆記試験を大別すると、客観式テストと論述式テストに分けられる。客観式テストに関しては多くの心理学的および教育学的な研究が行われてきたが、論述式テストの研究は大きく立ち遅れており、方法や問題点が十分明らかにされないままに実際の大学入学試験や大学内の試験に使われているのが実情である（渡部ほか, 1988)。

論述式のテストは、総合的な学力の評価に適した試験の形式である、と主張されることがある（北尾, 1991)。ここで、総合的学力とは、基礎知識の習得、適切な問題の設定、その知識に基づいて問題を考察する能力、解答を明瞭に表現する能力、などから定義される。論述式テストの好ましい特性に関するこのような主張は、人々の常識に近いからであろうか、多くの教育心理学の入門書に見られる。ところが、この主張を支持するデータは乏しく、多くの実証的データはこの主張を支持しない。支持しない研究の論点は二つある。一つ目は小論文の成績は、課題により同じ解答者でも異なるという課題による制限であり、これは測定に使う道具が対象となる能力を正しく測定できないという妥当性の低さにつながる。二つ目は小論文の評価が評定者の間で一致しないという信頼性の低さである。評定者は小論文の得点に影響する要因である（Cooper, 1984；渡部ほか, 1988；渡部, 曹, 1992；梶井, 2002)。妥当性と信頼性が低ければ、能力や学力を測定する道具としては適当ではない。したがって、論述式テストは学力や能力を測定する道具として望ましいものではない、というのが論述式テストの心理学的研究から得られる一般的結論である（例えば、池田, 1992)。

理論と実践は異なるもので、大学における成績の評価は、今日でも論述式テストによることが多い。大学における成績の評価は、客観式テスト（いわゆる○×式）より論述式テストのほうが好ましいと思う大学の教員が多いのであろう。大学で学ぶ内容を、○×式のテストで評価することはできない、という主張には説得力がある。特定の科目（作文以外）の比較的限定されたテーマを使う論述式テストは、広く一般的テーマを使う作文・小論文とは、評価の観点が異なるであろう。特定科目の論述式テストでは、知識や論理的思考により重きが置かれ、これらの事項の評価はより客観的基準に近いかもしれない。このように考えると、大学における論述式テストを、大学入試の小論文テストや中学・高校の作文と同じように扱えるかは、疑問である。この疑問に答えるためには、実際に大学の教科で

＊岩手大学教育学部　＊＊岩手大学大学院教育学研究科

行う論述式テストの評価を研究する必要がある。そこで、本研究は、大学の教科のレポートを材料にして、評定者間で評定値がどの程度一致するかを評点者の要因を考慮して検討した。

レポートの評価１

方法

岩手大学教育学部の2004年度前期選択科目（１科目）の受講生のレポートを用いた。レポートの課題は、授業で学習した知識に基づいて、自分の意見をまとめるものであった。12人の学生のレポートを筆者たち４人が独立に30点満点で採点した。１人は岩手大学教育学部の教員、他の３人は教育学部研究科の大学院生である。採点に必要な基礎知識は、あらかじめ筆者の１人から他の３人に説明されたが、評定の観点や統一した評価の方針は定めなかった。４人の評定者は独立に12人のレポートを評定し、この４人の評定値を分析対象とした。

結果

評定者には個人差があった。Table 1 から、評定者間で平均値に差はないが、分散は評定者Ｂが他の評定者より小さいことがわかる（有意差あり、$p<0.05$）。ＡとＤ、ＢとＣが点数の分布において、類似している（Fig. 1）。ＢとＣはともに少数の得点しか使用しなかった。

Table 1　各評定者の平均値と分散

	評定者Ａ	評定者Ｂ	評定者Ｃ	評定者Ｄ
平均	17.83	20.42	24.17	17.00
分散	35.24	4.45	17.42	31.09

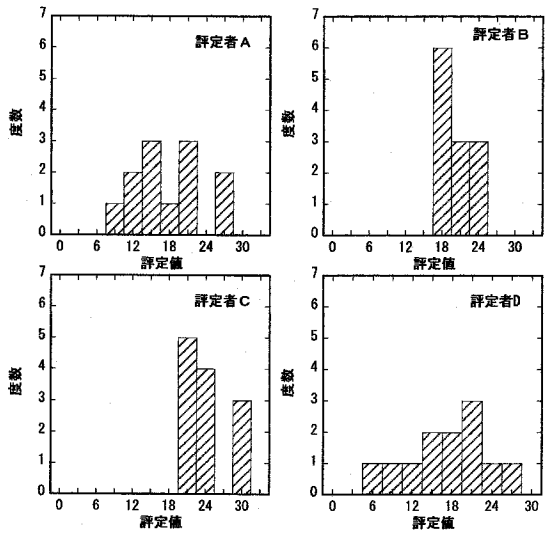

Fig. 1　レポート１に関する４人の評定者による評定値の分布

評定者間の相関係数（Pearson の r）の平均は0.4840，標準偏差は0.1269であった。評定者４人（Ａ，Ｂ，Ｃ，Ｄ）の間の点数の関連を散布図と相関係数で Fig. 2 に示す。評定者Ｂの点数は評定者Ｃ、Ｄの点数と有意な相関があった（$p<0.05$）が、評定者Ａの点数とＤの点数の相関は低かった。Fig. 2 に示す散布図のパターンは、評定者間の一致度が４人の評定者では異なり、不安定であることを示す。作文を複数の評定者が採点し，その相関を調べた先行研究は、0.40 (Coffman, 1966)、0.35～0.51 (安

藤, 1874)、0.70（池田, 1992）の値を見出している。レポート１の結果は、先行研究の結果にほぼ一致する。

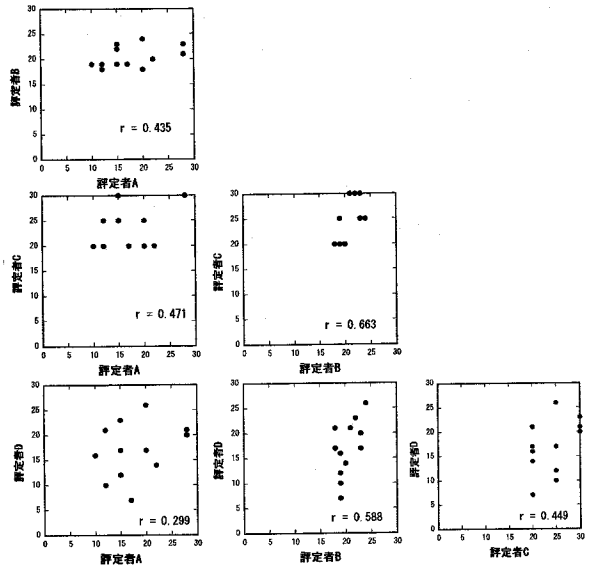

Fig. 2　レポート１に関する４人の評定値間の相関
評定者間で実質的な相関が低いことが読み取れる。

レポートの評価２

レポートの評価１の標本数は12と小さかったため、偶然の要因が強く働いた可能性がある。レポート２では、標本数を大きくした。加えて、評定者が評定値の範囲を有効に使えるようにするために、10点満点の採点方式にした。30点満点では、０から30までの31段階の評定は考えにくいが、10点満点の場合は、０から10までの11段階の評定が可能であろう。

方法

岩手大学教育学部の2004年度前期選択科目（１科目）の受講生38人の宿題のレポートを用いた。レポート２の科目はレポート１の科目と異なり、受講生も異なる。レポートの課題は、授業で学習した知識に基づいて、自分の意見をまとめるものであった。筆者たち４人が独立に10点満点で全てのレポートを採点した。４人の評定者はレポートの評価１と同じである。採点に必要な基礎知識は、あらかじめ第一筆者から他の３人に説明されたが、評定の観点や統一した評価の方針は定めなかった。４人の評定者は独立に38人のレポートを評定し、この４人の評定値を分析の対象とした。

結果

評定者間で個人差があった。Table 2 からわかるように、評定者間で平均値に差はないが、分散はＢがＡおよびＣより小さく、ＤがＣより小さかった（有意差あり、$p<0.05$）。ＢとＤが点数の分布において類似している（Fig. 3）。ともに、平均値を中心に左右対称で正規分布に近い得点の分布を示す。ＡとＣにはその傾向が見られない。予想どおりに、ＡからＤまで全評定者が広い範囲で得点を与えた。

評定者間の相関係数（Pearson の r）の平均は0.317，標準偏差は0.287であった（中央値は0.208）。

評定者4人（A, B, C, D）の間の点数の関連を散布図と相関係数でFig. 4に示す。評定者Aの点数は、評定者Cの点数と有意な相関があり（$p<0.05$）、評定者BとDの点数間にも有意な相関があった（$p<0.01$）が、他の組み合わせの相関は低かった。Fig. 4に示す散布図のパターンは、評定者間の一致度が4人の採点者では異なり、不安定であることを示す。レポート2から得られた相関係数は、先行研究の相関係数よりも小さかった。

Table 2　各評定者の平均値と分散

	評定者A	評定者B	評定者C	評定者D
平均	5.37	6.05	4.74	5.97
分散	4.18	2.38	6.31	2.57

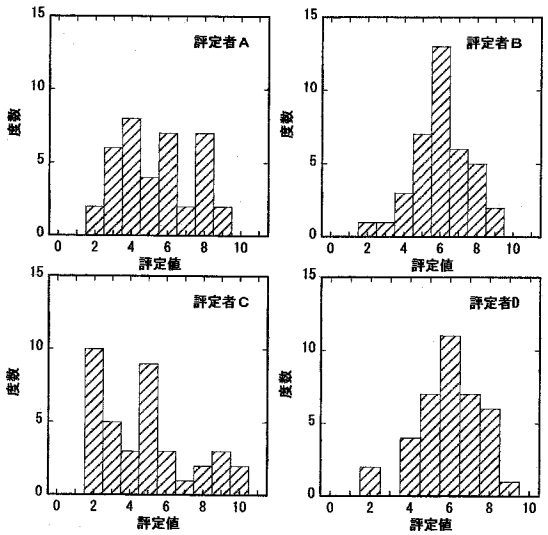

Fig. 3　レポート2に関する4人の評定者による評定値の分布
Fig. 1に比べると、どの評定者も評定尺度の広い範囲を使っていることがわかる。

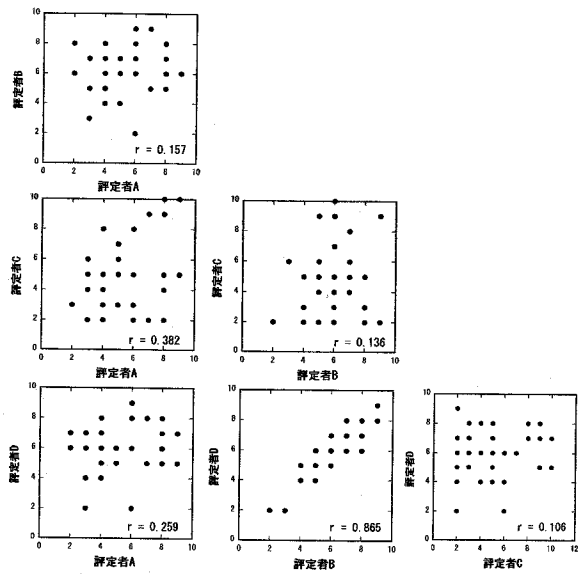

Fig. 4　レポート2に関する4人の評定値間の相関
1つの組み合わせ（評定者BとD）を除いて、評定者間で評定値に相関がない。

総合考察

二つのレポートに関しては授業科目と受験者が異なり、さらに標本数も異なるが、評定者の特性と評価の特性に関して一貫した結果が得られた。これらの特性は、これまでに報告された論述式テストや小論文テストに共通する特性である。

第一に、レポートの評価１と２を通して、評定者は比較的一貫した傾向を示した。評定者Ａは評定値の範囲を広く取り、点数の度数分布が一様分布に近い。これに対して、評定者Ｂは評定値の分散が小さく、点数の度数分布が山形分布に近い。評定者Ｃは低い点数で度数が高く、高い点数になるほど度数が減少する傾向を示した。評定者Ｄは一貫して点数の度数分布が山型分布に近かった。この個人内の一貫した評定の傾向は、同じ教員がレポートや学期末の論述式テストを採点するならば、複数のレポート課題、複数の論述式テストの評価は一貫する可能性を示唆する。

先行研究の中には、この個人内の要因を指摘しているものがある。例えば、手書き答案とワードプロセッサーを使って書いた答案に対する評定を比較すると、複数の評定者による評定値の平均に違いはないが、二つのタイプの答案に対する評定に評定者間で個人差があり、個人内で一貫した傾向があった（渡部, 曹, 1992）。また、作文の評価は、採点者が「作文が得意」と「作文が好き」（自分で作文を書くことが好き、ということ）という要因の影響を受け、文章を書くことを得意とする評定者と不得意と感じている評定者が、採点において異なる評価の基準を使っていた（梶井, 2002）。

第二に、レポートの評価１と２を通して、評定者間の一致度は低かった。全体で12の散布図が得られたが（Fig. 2 と Fig. 4）、その中で、レポート２の評定者Ｂと評定者Ｄの散布図だけが、評定者間の高い一致度（高い相関）を示し、ほかの散布図では評定者間の一致が見られない。例えば、レポートの評価１で評定者Ｂは評定者Ｃと高い相関（$r = 0.663$, $p < 0.02$）があったが、散布図（Fig. 2）で見ると、評定値が狭い範囲の特定の値に集まっており信頼できる一致度ではないことがわかる。レポート評価１の評定者Ｂと評定者Ｄの場合も同様である。レポート評価２の散布図は、評定者間で評定値が一致していないことをより明確に示す（評定者ＢとＤの場合を除く）。この評定者間で見られる傾向から、複数の採点者が分担してレポートや試験の一部分を採点すると、その評価は一貫しておらず、信頼できないと考えられる。

この評定者間の不一致は、作文の評定（例えば、Coffman, 1966；安藤, 1974；池田, 1992）に関しては繰り返し発見されてきた事実であり、歴史の記述試験においても見出されている（Swineford, 1964…池田, 1992からの引用）。しかし、筆者たちの知る範囲では、日本の大学における履修科目の課題レポートに関する研究は報告されていない。

他方、レポートの評価１と２の違いとして、30点満点と10点満点の違いが現われた。10点満点と範囲を小さくした場合は、予想どおり評定者は尺度全体を使う採点を行った。ただし、レポート評価１とレポート評価２の間では、標本の大きさが違うので、標本サイズの要因が交絡している。そのため、一義的な解釈はできない。

本研究では論述式テストの答案を用いる代わりに、宿題のレポートを評定した。評定に含まれる評定者の要因に関しては、論述式テストの答案もレポートも類似していると仮定したからである。しかし、特定の試験日に特定の時間制限の中で行う論述式のテストと、ある期間が与えられ参考資料を比較的自由に使える宿題のレポートを同じものとして扱えない側面もある。レポートはいろいろなものを使って調べて作成できる。例えば、インターネットを使って検索し、そのまま文章を添付したと疑われるレポートがいくつかあった。そこまで極端ではなくても、類似のレポートがいくつかあり、本

人が考察したか疑わしい例もあった。現実に大学の教員はレポートを評定して成績を決める資料にしているのであるから、レポートに特有の問題を考慮に入れて採点する方法を検討し、加えて、レポートの評定と論述式答案の評定を比較する研究が必要であろう。

【引用文献】

安藤　公平　1974. 小論文採点法の一検討　日本教育心理学会総会発表論文集、vol.16, pp.492-493.

Coffman, W.E.　1966. On the validity of essay tests of achievement. Journal of Educational Measurement, 3, (2), 151-156.

Cooper, P. L.　1984. The assessment of writing ability : a review of research. GRE Board Research Report GREB No.82-15R ETS Research Report 84-12.

池田　央　1992. テストの科学　日本文化科学社

梶井　芳明　2002. 児童の作文評価に影響を及ぼす評定者の心理的要因に関する研究　電子情報通信学会技術研究報告　TL. 思考と言語　vol 1, 102 Num. 491, 25-30.

北尾　倫彦　1991.「第4章学習の評価」教育心理学［新版］（北尾、杉村、山内、梶田　著）有斐閣

渡部洋, 曹亦薇　1992. 小論文評価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　第32巻

渡部洋, 平由美子, 井上俊哉　1988. 小論文評価データの解析　東京大学教育学部紀要　第28巻