論述式テストの研究(1)：採点者間の一致度

阿久津洋巳*　菊池梢**　鈴木安澄**　鈴木光**　渡邉愛枝**

(2006年2月6日受理)

Hiromi AKUTSU, Kozue KIKUCHI, Azumi SUZUKI, Hikaru SUZUKI and Yasue WATANABE

A Study of Essay Tests (1)：Agreement between Raters

はじめに

論述式テストは，客観式テストが測定できない思考能力を測定できると一般に理解されている。この思考能力には，分析的思考，批判的思考，問題発見能力，問題解決能力などが含まれる。だが，論述式テストがどのような思考能力をどの程度測定できるのか，どのようにすれば信頼できる測定ができるのかなどといった，実施に必要な条件を見出せるほどに心理学的および教育学的な研究は進んでいない（平井，2002ａ）。適切な実施方法や問題点が十分解明されないまま，様々な試験に使われており（渡部・平・井上，1988），現在もこの状態に変わりはない。本研究は論述式テストの問題の中で重要な課題である論述式テストを採点する際の採点者の主観の影響について検討する。

採点者の主観が影響せず，誰が採点しても同じ採点が得られるならば，そのテストによる測定は信頼性が高いといえる。この点に関して論述式テストは，客観式テストに比べて大きく劣る。論述式テストの答案を採点する際に，その採点値が採点者間で変動する原因には，大きく分けて２つの要因が考えられる。

第一の要因は，問題および解答の性質による影響である。採点者間で変動が生じやすい問題や解答というものがある。例えば，知識を問う形式の問題は，意見を問う形式の問題に比べれば，採点者の影響は小さいであろう。長い答案によい点数を与える傾向が，多くの採点者にある（平井，2002ｂ）と仮定すると，字数制限により記述量を統制する場合は，統制しない場合よりも，採点者間で採点値のばらつきが大きくなるであろう。また，事前に問題が提示され受験者が解答を準備できるなど，使用する知識量を統制しない場合では，そうでない場合に比べて採点者の影響の程度は変わる可能性がある。他に，答えさせる方向と内容を分析して大まかに制限を加えた問題を作った場合，採点者間で変動が生じにくいと予想できる（長澤，2003)。例えば，“不登校の改善策について意見を述べよ”という問題の場合，不登校を肯定する立場の解答を制限できると考えられる。

第二の要因は，採点者側の影響である。各採点者が異なる観点から答案を採点することがある。例えば，採点者間でいくつかの，暗黙あるいは明示された，採点項目のうち重要性のおき方が異なったり，答案の解釈が異なったりする。採点者の主観が採点に影響しにくい方法としては，評価する項目を多数列挙して，項目ごとの配点を設定する分析的評価が挙げられるが，分析的評価を行うには時間と労力がかかる。また，分析的評価の結果が総合的評価より信頼性が高いわけではなく（渡部ほか，1988），多くの採点者が直感的に良

＊岩手大学教育学部　＊＊岩手大学大学院教育学研究科

いと感じた答案が高い得点を得られない（池田，1992），などの問題点が指摘されている。そこで，評価の方法に多少の制限を加える方法，例えば，明瞭な基準を用意して，それに沿って採点する方法，知識の程度や理解の程度など，確認しようとする目標に観点を絞って採点する方法などが考案されている（長澤，2003）。

第一の要因（問題および解答の性質による影響）は，第二の要因（採点者側の影響）の効果を制限するであろう。論述式テストの解答を採点する際，採点者の主観が影響するとしても，あらゆる論述式問題で一様にその影響があるとは考えがたい。例えば，受験者の能力の影響が最大要因と仮定すると，試験を受ける集団が，解答の質および量に大きなばらつきを示すような問題では，答案がどれも似通っている場合に比べ，採点者以外の影響が大きいであろうと考えられる。

以上のことから，採点者間で採点値に変動が比較的生じにくい論述式テストの条件として，正答と誤答がはっきりしている問題や，採点者の主観の相違よりも答案のでき不出来が大きく目に見えるような問題が考えられる。このような問題を用い，さらに，採点の基準を設定すれば，採点者間の採点が一致しやすく，信頼性が高いと予想される。日本の教育評価の教科書に述べられている方法(例えば，長澤，2003)は，論述式テストによって測定される能力と測定法ともに理論的な予測であり，これらの方法の正当性を証明する実証的研究はほとんど報告されていない（平井・椎名・柳井，2001)。理論を十分な実証的検討を経ずに、実用の指針として主張したり，現実に適用することは，科学知識の応用としては望ましくない。

そこで，本研究は，論述式テストの問題の中で重要な課題である論述式テストを採点する際の採点者の主観の影響について明らかにすることを目的とする。具体的には，採点者間で比較的採点が一致しやすいと思われる問題を用い，採点の基準を設ける条件と設けない条件とを比較し，採点者間で生じる採点値の変動を制限する方法を実験的に検討する。

実験 1

目的

採点者間で採点値に変動が生じにくいと思われる論述式問題に対する答案を複数の採点者が採点し　その一致度を調べた。一致度を高めるために，いくつかの採点基準を与えた。

方法

岩手大学全学共通教育科目の選択科目（１科目）の受講生から論述式テストの答案を得た。論述式の試験課題は，講義で学習した知識に基づいて解答する形式であり，２つの項目を含んでいた。事前に課題を提示し，試験当日，受験者が解答を制限時間内で教科書その他の資料を参照せずに記述する形式である。全受験者27名の答案を岩手大学大学院教育学研究科の大学院生４名（以下Ａ，Ｂ，Ｃ，Ｄと呼ぶ）が採点した。採点者４名は，独立に10点満点で全てのレポートを採点した。採点者には，採点に必要な科目内容の知識と明確な採点基準を与えた。採点基準は，平均点の目安と配点方法である。具体的には，「①平均点を５点程度とする。②10点満点のうち，項目１に５点，項目２に５点と配点する。③300字以上の解答が好ましいので，それより短い解答は，一律に２点減点する。」という３つの基準である。採点者は，全員一室に集まり，1.5時間から２時間ですべての採点を完了した。採点作業中，採点者間で採点法や内容について話し合うことは許可されず，また，採点者は全員他の採点者の採点値を知ることはなかった。

結果

採点者間で採点平均値に大きな違いはなく（Table 1），分散分析の結果，平均値間に有意差はなかった（$F=1.91, df\ (3,104)$，$p>0.132$）。また，各採点者の採点平均値は，目標としていた平均値５と有意差がなかった（平均値ごとのｔ検定による。$p>0.07$）。

採点値の分布については，ＡとＤ，ＢとＣが類

似している（Fig.1）。加えて，すべての採点者において，採点値が広い範囲に分布するとともに中心化傾向が見られた。

Table 1 各採点者の採点平均値と標準偏差（実験1）

採点者A	採点者B	採点者C	採点者D
4.59	4.59	4.78	5.78
(2.32)	(1.78)	(2.24)	(2.15)

（ ）内は標準偏差

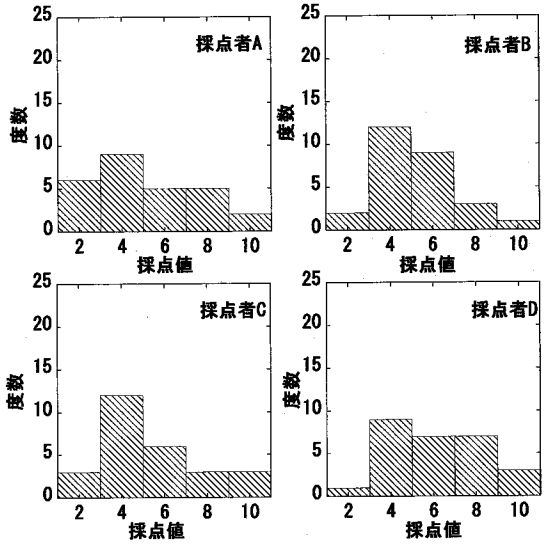

Figure 1　4人の採点者による採点値の分布（実験1）

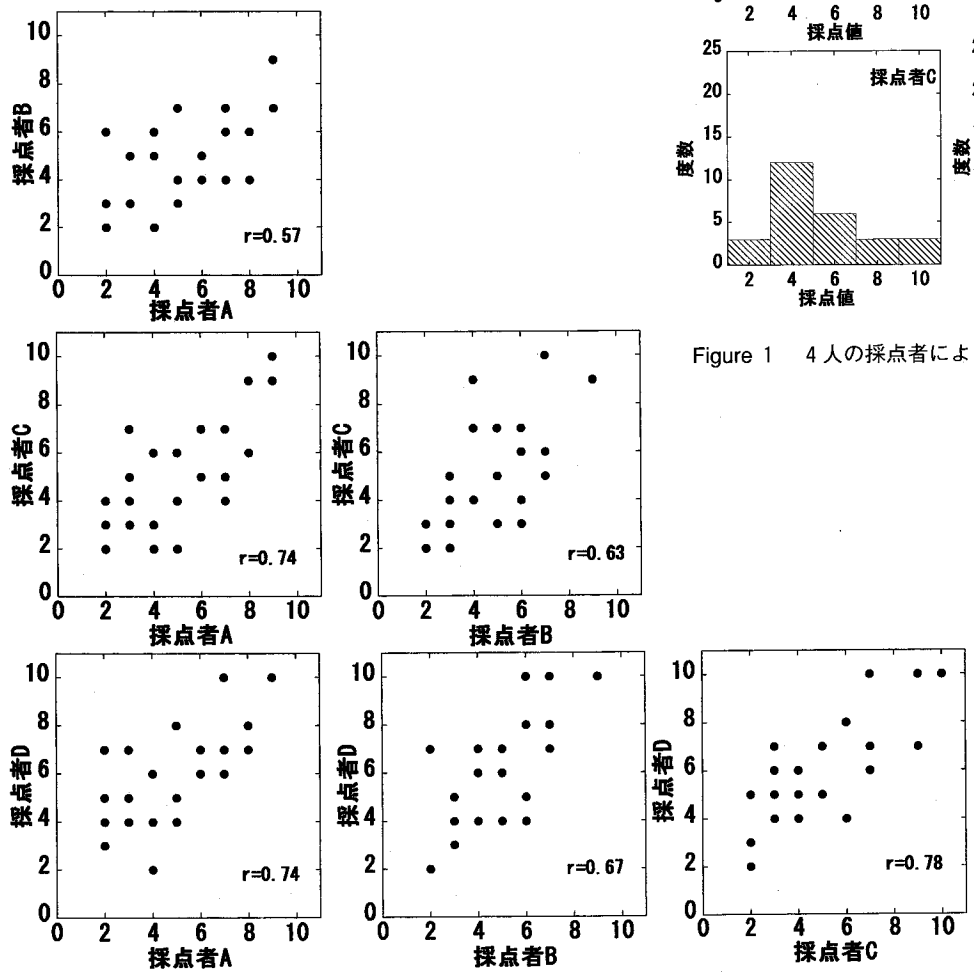

Figure 2　4人の採点値間の相関（実験1）

採点者間で相関が高い。

採点者間の相関係数(Pearson の r)の平均は0.69，標準偏差は0.08であった。採点者間（A～D）の採点値の関連を散布図と相関係数で Fig.2に示す。全ての採点者間において，0.57～0.78の強い正の相関が認められた。作文を複数の採点者が採点をし，その相関を調べた先行研究では，0.40（Coffman,1966），0.35～0.51(安藤，1974)，0.26～0.43（渡部ほか，1988)，0.70(池田，1992)0.32～0.48（阿久津・嶋野・熊谷・佐々木，2005）の値を見出している。本研究の実験1における相関係数の値は，池田（1992）を別にすれば，これらの先行研究よりも高い傾向にある。

実験2

目的

採点者間で採点値に変動が生じにくいと思われる論述式問題に対する答案を複数の採点者が採点

し，その一致度を調べた。実験１の結果と比較して採点基準の効果を調べるために，実験２では採点基準を与えなかった。

方法

答案は，実験１と同じものを使用した。採点者は，岩手大学大学院教育学研究科の大学院生４名（以下E，F，G，Hと呼ぶ）である。採点者は，全員実験１の採点者とは異なる。採点者４名は，独立に10点満点で全てのレポートを採点した。採点者は０～10点の範囲で採点するようにだけ指示され，採点基準は説明されなかったが，採点に必要な問題内容に関する知識は，実験１と同様に与えられた。実験１と同じく，採点者は，全員一室に集まり，0.75時間から1.5時間ですべての採点を完了した。採点作業中，採点者間で採点法や内容について話し合うことは許可されず，また，採点者は全員他の採点者の採点値を知ることはなかった。

結果

採点者間で採点平均値に差異がみられ(Table 2)，分散分析によって統計的有意差が確かめられた（$F=30.92, df(3,104)$，$p<0.0001$）。TukeyのＨＳＤを用いた多重比較によれば，採点者Ｅ，Ｆ間以外に有意差があり，平均の大小関係は「採点者Ｅ＝Ｆ＞Ｈ＞Ｇ」であった。また，採点者Ｅ，Ｆ，Ｇの採点平均値は，実験１で期待した平均値５と有意差があった（平均値ごとのt検定による。$p<0.0001$）。

採点者Ｅ，Ｆ，Ｇ，Ｈによる採点値の分布を調べると，採点者Ｆ，Ｇは採点値に大きく偏りがあった（Fig.3）。１名の採点者（Ｈ）は，実験１の採点者と同じような採点値の分布を示した。評価基準が与えられなかったので，各採点者独自の配点の傾向が現われたのであろう。

Table 2 各採点者の平均値と標準偏差（実験2）

採点者E	採点者F	採点者G	採点者H
6.85	7.22	3.93	5.52
(1.66)	(0.75)	(1.17)	(1.76)

（　）内は標準偏差

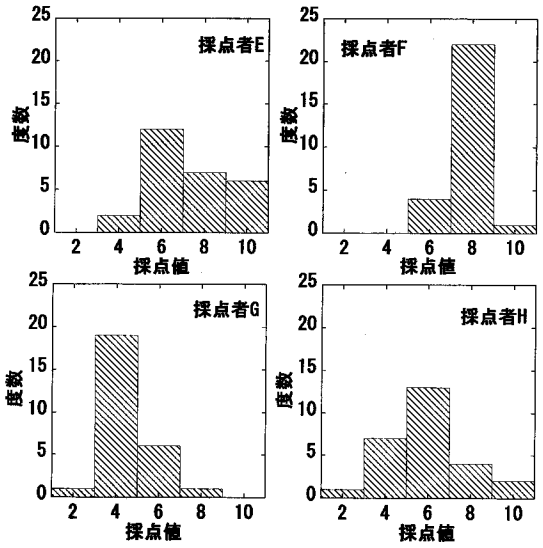

Figure 3　４人の採点者による採点値の分布（実験２）

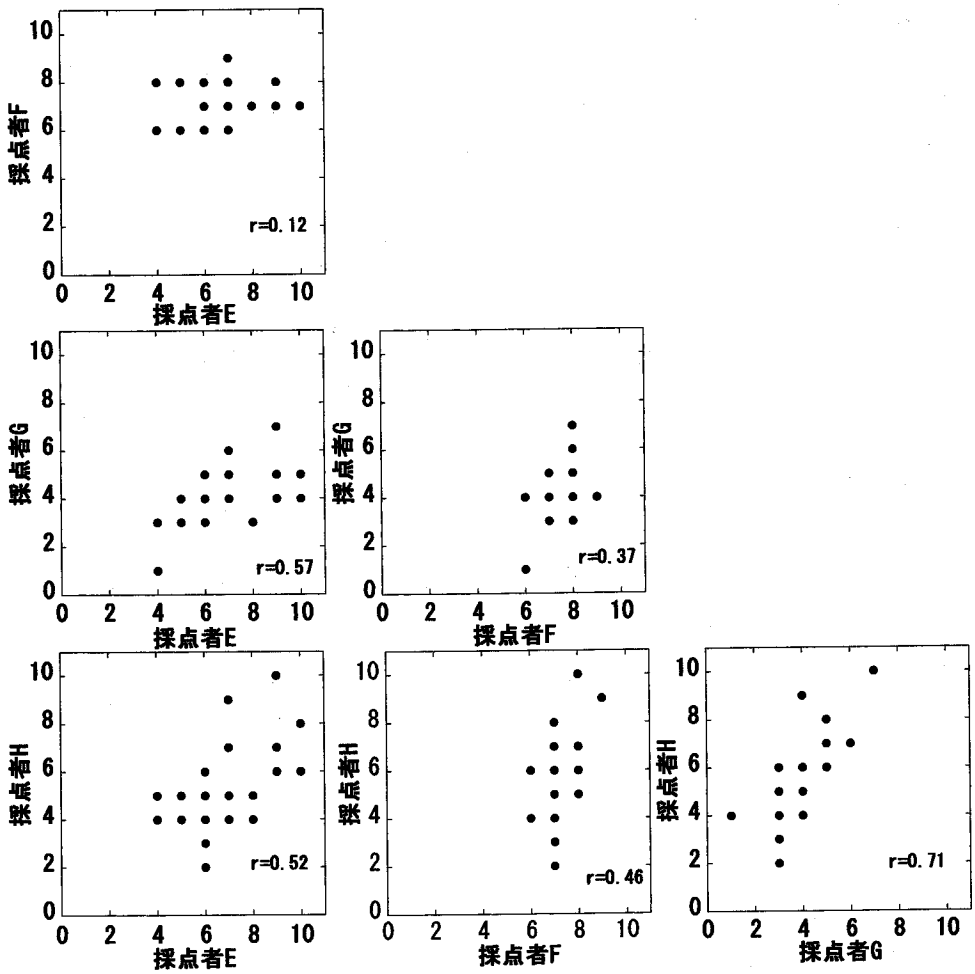

Figure 4　4人の採点値間の相関（実験1）　　採点者間で相関が低い。

採点者間の相関係数(Pearsonのr)の平均は0.46，標準偏差は0.20であった。採点者間（E～H）の採点値の関連を散布図と相関係数でFig.4に示す。採点者G－H間では，強い正の相関があり，採点者E－G間，E－H間，F－H間では比較的強い正の相関があった。採点者F－G間では弱い正の相関があった。また，採点者E－F間では，ほとんど相関はなかった。相関係数の平均値0.46は，先行研究で報告されている相関係数の平均値に近い（実験1の結果を参照）が，実験1の平均値0.69よりも有意に小さい値であった（$t=2.647$，$df=6.52$，$p<0.017$）。

総合考察

本研究の主目的は，採点者側の影響を調べることであるので，第二の要因から考察する。

第二の要因について

知識を問う形式の問題を用いたにも関わらず，実験1と実験2における採点者間の相関は異なる結果となった。採点者間の相関係数を実験間で比較すると，実験1よりも実験2の方が採点者間の相関が低く，採点者間の相関係数の平均値は実験1では0.69，実験2では0.47であった（有意差あり，$p<0.017$）。さらに，実験1よりも実験2の方が採点者間の相関係数にばらつきがみられ，採点者間の相関係数の標準偏差は，実験1では0.08，実験2では0.20であった（有意差あり．$F=6.404, df(5,5), p<0.032$）。問題および解答の影響ではなく，採点者側の影響が，実験1および2の結果に現れた。実験1では採点者間の一致度が高く，採点の信頼度が高かった。

実験1は3つの採点基準を設定したことが実験2と異なった。これらの基準について以下に検討

する。

採点基準①：平均点を５点程度とする

結果に示したように，実験１では採点者全員が期待された平均値に近い値となるように採点したが，実験２では１名を除いて期待された平均値から離れた値となった。この平均値の設定は，採点後に統計的に採点者間で平均値を一定にすることとは別である。平均値を設定することで，各採点者が５点を中心としてその上下に採点値を配分するため，採点値の分布が採点者間で近く（Fig.1参照），採点値の分布が，採点者間で一致しやすくなる効果がある。採点値をバランスよく配分したため，「全員に高い得点を与える」「全員に低い得点を与える」といった極端な採点が生じなかったともいえる。

採点基準②：10点満点のうち，項目１に５点，項目２に５点と配点する

実験１では共通の配点条件があったのに対し，実験２では各採点者が独自の配点法により採点したと考えられる。採点者間で採点値の一致度が低下した原因のひとつであろう。採点の際，１つの問題を２つの独立した項目に分けるという方法は，分析的評価法に近い。問題を項目分けせずに全体で10点という配点法を用いた実験２に比べると，実験１は各項目５点の配点であるから，採点者間の採点値の変動が小さくなる。加えて，多くの項目について独立して採点する場合に比べ，労力も時間も節約できる。今後，試験の目的に応じて項目数をいくつにするのが適当かを検討する必要がある。

採点基準③：300字以上の解答が好ましいので，それより短い解答は，一律に２点減点する

該当した答案の採点最高値は８になり，最高値10の場合に比べると，若干ではあるが，採点者間で与える採点値の範囲が限定され，より一致しやすかった。しかし，字数の基準で減点された答案は３枚であり，その影響は基準①②に比べると小さな要因だと考えられる。

第一の要因について

次に試験問題の特性について検討する。この実験に参加した採点者８名は，実験１および２の約６ヶ月前に，別の講義の課題レポートの採点を行った（以下実験０と呼ぶ）。実験０に用いたレポートの課題は，講義で学習した知識を利用して，自分の意見をまとめるものであった。受講者８名のレポートを，実験１および２と同じ採点者（A～H）が独立に20点満点で採点した。この際，採点に必要な知識は説明されたが，採点の基準や，配点の方法は定めなかった。実験１の採点者A，B，C，D（以下採点者群Ⅰと呼ぶ）の採点値の相関係数（Pearson の r）は，範囲が（−0.280, 0.534）平均が0.127，標準偏差が0.270であり，実験２の採点者E，F，G，H（以下採点者群Ⅱと呼ぶ）の採点値の相関係数は，範囲が（0.104，0.802）平均が0.303，標準偏差が0.251であった。採点者群ⅠおよびⅡの間で，相関係数の平均値に有意差はなかった（t＝−1.168，df＝9.946，p＞0.134）。（付録に各採点者の平均採点値と標準偏差を２群に分けてTable 3，Table 4として示した。採点者間の相関も２群に分けてTable 5，Table 6として示した。）

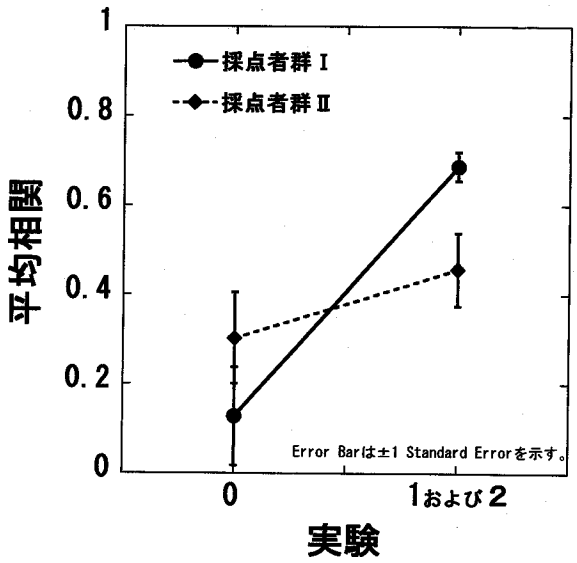

Figure 5　採点者群内の相関を実験０と実験１および２について示す。
実験０では２つの採点者群間に違いはないが、実験１および２では採点者群Ⅰは採点者群Ⅱより相関が高かった

採点者群Ⅰは，実験0および実験1の間で平均相関係数が大きく変化したが，採点者群Ⅱは実験0および実験2の間で，その変化が小さかった（Fig.5）。実験0と実験1および2，さらに2つの採点者群を実験要因とみなして，1要因で繰り返しがある2元配置の分散分析を行ったところ，実験の主効果と採点者群×実験の交互作用が有意であったが（実験の主効果は，$F=18.32, df\ (1,10)$，$p<0.002$；交互作用は，$F=5.76$，$df\ (1,10)$，$p<0.037$），採点者群は有意ではなかった（$F<1.0$）。TukeyのＨＳＤを使って平均相関係数間の有意差を求めると，採点者群Ⅰでは実験0および1の間に有意差があった（$p<0.05$）が，採点者群Ⅱでは，実験0および2の間に有意差はなかった（$p>0.05$）。実験0では，採点者群ⅠおよびⅡの間に有意差はなく，実験1および2では，両群間に有意差があった（$p<0.05$）。

ここで，採点者群ⅠとⅡの間では各人の採点傾向にもともと違いがあり，採点者群ⅠはⅡよりも似通った採点をする人が集まったという解釈もあるかもしれない。しかし，実験0の結果は，2つの群が採点の一致度について違いがないことを示し，2つの採点者群は同質といえる。

意外なことに，新しい採点条件が与えられなかった採点者群Ⅱにおいては実験0と実験2の間で採点者間の一致度はさほど大きく変化しなかった。これに対し，採点条件が与えられた採点者群Ⅰにおいては実験0と実験2の採点結果の一致度に違いが見られた。このことから，実験1および2では，第一の要因（問題および解答の性質による影響）を考慮して採点者間で採点値に変動が生じにくいと考えられる論述式問題を用いたものの，十分にその意図に沿った問題ではなかったといえる。この点は今後の検討が必要である。

今後の展望

本研究では，比較的少数の基準を用いることで，採点者間で採点値の一致度が高くなることが示された。今後は，思考能力に重点を置く論述式テストの採点に適用して検討していく必要がある。前述したように，正答と誤答がはっきりしている問題や，採点者の主観の相違よりも答案の出来不出来が大きく目にみえるような問題の場合には，採点者間の採点値の変動は比較的小さいと考えられる。しかし，それに伴い，論述式テストの利点である思考能力の測定から離れてしまう可能性がある。論述式テストを用いることにより，客観式の学力テストで測りにくいと思われる情報理解力，視野の広さ，発想の展開力，要約力などが測定可能であると報告されている（平井ほか，2001）。この論述式テストの本来の利点を活かすためには，採点者間の変動を生み出す第一の要因（問題および解答の影響）を制限せず，第二の要因（採点者側の影響）を統制する方が望ましいであろう。

付録

実験0のデータを表にして以下に示す。採点者は実験1および2と同一採点者である。

Table 3 各採点者の採点平均値と標準偏差(実験0)

採点者A	採点者B	採点者C	採点者D
15	14.75	15.63	15.38
(1.60)	(1.67)	(2.33)	(2.83)

(　)内は標準偏差

Table 4 各採点者の採点平均値と標準偏差(実験0)

採点者E	採点者F	採点者G	採点者H
15.75	16.25	13.75	14.25
(1.98)	(1.04)	(1.75)	(4.27)

(　)内は標準偏差

Table 5 各採点者間の採点値の相関（採点者群Ⅰ 実験0）

	採点者A	採点者B	採点者C
採点者B	0.534		
採点者C	0.230	0.193	
採点者D	0.095	−0.008	−0.180

Table 6 各採点者間の採点値の相関（採点者群Ⅱ 実験0）

	採点者E	採点者F	採点者G
採点者F	0.104		
採点者G	0.802	0.276	
採点者H	0.194	0.243	0.201

謝辞

本研究の実験実施に当たり，採点者として協力していただいた岩手大学大学院教育学部研究科の学生に感謝します。

引用文献

阿久津洋巳・嶋野恵美子・熊谷賢・佐々木和歌 2005　課題レポート評価における評定者間の一致　岩手大学教育学部附属教育実践総合センター研究紀要　第4号, 75－80.

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.

平井洋子・椎名久美子・柳井晴夫　2001　文系学生向き総合論述問題による能力測定の試み　大学入試センター研究紀要　No 30，1－20.

平井洋子　2002a　総合論述式課題の予備調査について　大学入試センター研究会開発部研究成果報告書「大学入学者選抜資料としての総合試験の開発的研究」所収

平井洋子　2002b　論述式課題による高次思考能力測定の試み　東京都立大学名文学報第326号，17－30.

池田　央　1992　テストの科学　日本文化科学社

長澤俊幸　2003　辰野・石田編　2003年度改訂版教育評価法概説　第4章　図書文化

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