様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月　 ７ 日現在

機関番号：１７１０２
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１３２９８

研究課題名（和文）評定者及び評価基準が記述式テストの結果に及ぼす影響

研究課題名（英文）The effect of raters and scoring rubric on the score of descriptive test.

研究代表者
安永　和央（YASUNAGA, Kazuhiro）

九州大学・アドミッションセンター・准教授

研究者番号：８０７７７６６５

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究は，高校2年生約300名に実施された国語の大学入試問題の回答を対象に，評価基準や評定者が記述式テストの結果に及ぼす影響について検討を行った。評価基準の検討においては，評価基準A：正答，誤答，無回答，評価基準B：評価基準Aの内容に準正答を加えた評価を設定し，2つの評価基準を用いて評価を行った。その結果，特定の設問においては識別力が高くなることが示唆された。評定者の検討においては，3名の評定者の評価がどの程度一致しているかを検討した。その結果，評価をする際の判断が評定者に委ねられる設問においては，一致度が中程度にとどまり，評定者によって受験者の評価が大きく異なり得ることが示唆された。

研究成果の学術的意義や社会的意義
本研究成果の学術的意義は，わが国においてほとんど検討されてこなかった国語テストにおける記述式問題を対象に，評価基準や評定者がテストの結果に及ぼす影響を明らかにした点である。また，教育現場の協力により得られた成果であるため，実際に教育へ応用できる点が本研究成果の社会的意義である。

研究成果の概要（英文）：This study examines if variations in the scoring rubric and raters for a descriptive item affect test-takers’ evaluations in a Japanese language comprehension test. The test was administered to about 300 students in the second year of high school. Two types of scoring rubric were used. Rubric A provides the correct answer, the incorrect answer, and no-response. Rubric B had a gradual evaluation in addition to the contents of rubric A. The items were evaluated based on rubrics A and B. Therefore, rubric B, having a gradual evaluation, allowed us to comprehend which statements were missing compared to the correct answer. Moreover, rubric B increased the discrimination value in a certain item. Regarding the raters, the degree of agreement between the three raters’ evaluations was examined. Consequently, the level of agreement was moderate for items in which the decision to evaluate was left to the raters, suggesting that the evaluation could vary greatly depending on the raters.

研究分野：教育測定学

キーワード：国語テスト　大学入試　記述式問題　評定者　評価基準　一致度　字数制限　項目分析

様　式　Ｃ－１９，Ｆ－１９－１，Ｚ－１９（共通）

１．研究開始当初の背景

2020 年から大学入試センター試験に代わり，大学入学共通テストが始まった。当初，大学入学共通テストにおける主な変更点として，「思考力・判断力・表現力」を重視した評価が求められることから，大学入試センター試験では出題されなかった記述式問題が導入されることが予定されていた（文部科学省，2017）。

記述式問題においては，問題構成，内容や回答（解答）字数等，その出題に関する課題が挙げられているが，採点（評価）における公平性も大きな課題となる。多枝選択式問題の場合，正解の記号が決められていれば，どの評定者でもその記号を回答しているものを正答とし，それ以外を誤答として採点できる。一方，記述式問題においては，受験者の回答が多様となるため，評価基準の設定の仕方や評定者により評価が異なることも考えられる。しかしながら，大学入学者選抜試験においては，評価基準や評定者が採点に及ぼす影響について実証的な検討がほとんどなされないまま，記述式問題の導入が検討されてきた。大学入学者選抜試験は個人の処遇に多大な影響を及ぼすテスト（high stakes test）であることを考慮すると，評価基準や評定者がテストの採点に及ぼす影響については，実証的に検討される必要があると考えられる。

２．研究の目的

国語の大学入試問題における記述式問題を対象に，評価基準（解答類型）や評定者がテストの評価（採点）に及ぼす影響について，以下 2 つの観点から検討を行う。

（1）評価基準の設定が能力評価に及ぼす影響に関する研究

評価基準の設定の仕方に着目した検討を行う。具体的には，正答，誤答，無回答から構成される評価基準とこれらに準正答となる基準を加えた評価基準を設定し，これらの評価基準の違いが記述式問題の結果に及ぼす影響について検討する。

（2）評定者間における評価の一致度に関する研究

複数名の評定者が評価基準に従って回答を評価し，これらの評価がどの程度一致しているかを検討する。また，一部の設問においては，回答欄の字数制限がある場合と字数制限がない場合で，回答を評価する評定者が能力評価にどのような影響を及ぼすかについても検討を行う。

３．研究の方法

（1）評価基準の設定が能力評価に及ぼす影響に関する研究

高校 2 年生約 300 名を対象に，大学入学者選抜試験の問題を基に作成された国語テスト（多枝選択式問題 6 問，記述式問題問 8 問，合計 14 問）を実施した。本研究で検討した設問は，設問 4a：人間の社会とゴリラの社会にどのような共通点があるかを説明する問題，設問 4b：人間の社会とゴリラの社会にどのような相違点があるかを説明する問題であり，両設問とも回答に字数制限のない問題であった。2 種類の評価基準（A，B）を設定して評定を行った。評価基準 A は，類型 1：「正答」（1 点），類型 9：「その他の回答（誤答）」（0 点），類型 0：「無回答」（0 点）を設定した。評価基準 B は，評価基準 A の類型に類型 2 と類型 3：「準正答」（0.5 点）を加えたものである。両設問とも，正答するためには 2 つの内容に着目した回答を記述することが求められる。2 つの内容のうちの 1 つしか記述されていない場合には，類型 2 か類型 3 の「準正答」に分類される。項目分析の手法を用いて得点率及び識別力，解答類型分類率を算出した。

（2）評定者間における評価の一致度に関する研究

（1）と同様の国語テストの回答を対象に，訓練を受けた 3 名の評定者が（1）の評価基準 B に従って評価を行った。分析では Fleiss の κ 係数を求め，3 名の評価がどの程度一致しているかを検討した。また，（1）で説明した設問 4a と設問 4b では，A 条件：字数制限あり，B 条件：字数制限なしの回答欄を設定し，字数制限がある場合と字数制限がない場合の評定者間の一致度について検討した。具体的には，設問 4a では，A 条件：50 字以内で説明せよ，B 条件：字数制限なし，設問 4b では，A 条件：70 字以内で説明せよ，B 条件：字数制限なしの回答欄を設定した。

４．研究成果

（1）評価基準の設定が能力評価に及ぼす影響に関する研究

各設問の得点率及び解答類型分類率を表 1 に示す。まず，設問 4a の評価基準 A では，約 26%の回答が類型 9：「その他の回答（誤答）」に分類された。他方，評価基準 B では，約 18%の回答が，類型 3：「準正答」に分類されていた。この結果から，評価基準 A では誤答に分類されていた多くの回答が，正答の内容の半分に言及できているということを把握することができた。次に，設問 4b の評価基準 A では，約 90%の回答が類型 9：「その他の回答（誤答）」に分類され，類型 1：「正答」に分類される回答が少ないことから，得点率が低くなっていた。それに対して，評価基準 B では，評価基準 A で誤答に分類された回答の約 67%が類型 3：「準正答」に分類された。このことから，半数以上の受験者が，実際には正答の半分の内容は記述できているということがわかった。また，設問 4a と設問 4b ともに，準正答には部分点として 0.5 点が配点されているこ

とから，評価基準 B の方が，得点率が高くなった。設問 4b においては，評価基準 B の方が，識別力の値が高くなった。設問 4b の評価基準 A では，多くの受験者がその他の回答（誤答）に分類されることで，得点率も低く難しい設問となった。このような設問の場合は，評価基準 B のように類型を段階的に設定し，部分点を設けることで，受験者をより明確に弁別できる可能性が考えられた。

以上から，誤答をまとめて評価するよりも，回答を段階的に評価することで，正答の回答と比べてどのような記述が足りないかを把握することができ，設問 4b においては識別力が高くなることが示唆された。

表 1　設問4aと設問4bにおける項目分析の結果

設問	*N*	条件	解答類型分類率					得点率 ≪難易度≫	I-T相関 ≪識別力≫
			0[a)]	1	2	3	9[b)]		
4a	147	A	.007	.735	—	—	.259	.735 (.441)	.121 [-.041, .278]
		B	.007	.735	.014	.184	.061	.833 (.299)	.065 [-.098, .225]
4b	145	A	.014	.083	—	—	.903	.083 (.276)	.059 [-.105, .220]
		B	.014	.083	.034	.669	.200	.434 (.264)	.209 [.047, .359]

[a)] 無回答，[b)] その他の回答，( )内の数字はSD，[ ]内の数字は95%信頼区間

（2）評定者間における評価の一致度に関する研究

まず，各設問における 3 名の評価の一致度を表 2 に示す。κ 係数の解釈基準は，0.00 ~ 0.20：わずかに一致，0.21 ~ 0.40：まずまずの一致，0.41 ~ 0.60：中等度の一致，0.61 ~ 0.80：かなりの一致，0.81 ~ 1.0：ほぼ完全，完全一致であった。設問 1a，1b，1c，3，8a，8b は，多枝選択式問題であり，κ 係数は，0.91 〜1.00 の値であった。多枝選択式問題の評価は完全に一致すると予想されたが，6 問中 5 問において評価にわずかなズレが生じていた。これは，評価する際の記載ミスや評価結果をエクセルファイルに打ち込む際の入力ミスが原因だと推察される。設問 6 は，正答が一つに定まる記述式問題であった。この設問の κ 係数は 1.00 であった。正答が一つに定まる記述式問題では，評価が完全に一致していた。多枝選択式問題では，選択枝の数（例えば，「ア」に回答，「イ」に回答，「ウ」に回答，「エ」に回答），選択枝以外の回答（誤答），無回答と評価基準の段階数が多くなるが，正答が一つに定まる記述式問題では，正答，誤答，無回答の 3 つとなるため，多枝選択式問題に比べてミスが生じにくかったと考えられる。設問 2，5，7，9a，9b は，記述式問題であり，κ 係数は，0.53〜0.77 の値であった。これらの評価は，中程度の一致からかなりの一致まで様々であった。回答を分類する際に，正答に求められる条件が複雑ではなく，特定の内容が書かれているかどうかを判断する設問では，かなりの一致を示した。他方，正答に求められる条件が複雑で，かつ，注意点における判断が評定者に委ねられる設問では，一致度の値が中程度となった。

表2　各設問における評定者の一致度

設問	設問形式	κ係数	*p*	*N*
1a	選択式	.914	< .001	303
1b	選択式	.986	< .001	303
1c	選択式	.992	< .001	303
2	記述式	.531	< .001	303
3	選択式	.978	< .001	303
5	記述式	.620	< .001	303
6	記述式	1.000	< .001	306
7	記述式	.759	< .001	306
8a	選択式	1.000	< .001	306
8b	選択式	.983	< .001	306
9a	記述式	.703	< .001	306
9b	記述式	.766	< .001	306

以上から，多枝選択式問題のような，正答が明確な設問においても評価にズレが生じることがあり得ることがわかった。また，記述式問題では，特定の内容が書かれているかを判断する設問であれば，評価はかなりの一致を示したが，完全一致には程遠いことも明らかになった。さらに，「評価が評定者に委ねられる設問」においては，一致度が中程度にとどまり，評定者によって受験者の評価が大きく異なり得ることが示唆された。

次に，A 条件：字数制限あり，B 条件：字数制限なしの回答欄を設定した設問 4a と設問 4b における 3 名の評価の一致度を表 3 に示す。設問 4a の κ 係数は A 条件が 0.67，B 条件が 0.81 であった。設問 4a における A 条件の評価はかなりの一致を示したが，B 条件の一致度の値は A 条件よりもさらに高く，ほぼ一致していた。設問 4b の κ 係数は A 条件が 0.65，B 条件が 0.72 であった。設問 4b では，A 条件，B 条件ともにかなりの一致を示したが，B 条件の方が A 条件よりも値が高かった。回答欄に字数制限がない場合，受験者の回答字数は少ないものから多いものまで幅広くなる。字数が多い回答を評価する場合，評定者が判断すべき箇所が増えるため，字数制限が定められている場合よりも評価の一致度が低くなることが予想される。しかし，本研究の結果はこの予想と異なり，字数制限がある条件よりも字数制限がない条件の方が一致度の値は高くなった。その原因の一つとして，字数制限がない条件の方が，誤答の判断がしやすかった可能性が考えられる。

表3　設問4aと設問4bにおける評定者の一致度

設問	設問形式	κ係数	*p*	*N*
4a: A条件	記述式	.669	< .001	155
4a: B条件	記述式	.808	< .001	148
4b: A条件	記述式	.649	< .001	157
4b: B条件	記述式	.716	< .001	146

今後は，回答内容と評価における一致度の関係について詳細に検討する。さらに，評価基準の段階数に関する知見を蓄積していくとともに，評定者が多数いる状況を検討し，分析結果や評定者へのインタビューを通して，評価の一致度を促進する要因についても検討を行う予定である。

＜引用文献＞

①文部科学省 (2017). 大学入学共通テスト実施方針　Retrieved from https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2017/10/24/1391397_001.pdf (2021 年 6 月 7 日)

5．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 安永和央・野口裕之
2．発表標題 評価者間における評価の一致度に関する検討－国語の大学入試問題を用いて－
3．学会等名 日本心理学会第84回大会
4．発表年 2020年

1．発表者名 安永和央・野口裕之
2．発表標題 国語の記述式問題における評価基準の検討－評価基準の段階数に着目して－
3．学会等名 日本心理学会第83回大会
4．発表年 2019年

1．発表者名 安永和央・野口裕之
2．発表標題 記述式問題における字数制限が回答に及ぼす影響－国語の大学入試問題を用いた実証研究－
3．学会等名 日本心理学会第82回大会
4．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関