Japanese-Language Articles and Reviews

日本の高校におけるスピーキング評価の採点者信頼性―教室内グループ型のディスカッションとディベートの場合

Rater Reliability in Speaking Assessment in a Japanese Senior High School: Case of Classroom Group Discussion and Debate

小泉利恵
Rie Koizumi
清泉女子大学
Seisen University

初澤晋
Susumu Hatsuzawa
宮城県石巻高校
Ishinomaki High School, Miyagi

磯部礼奈
Reina Isobe
宮城県石巻高校
Ishinomaki High School, Miyagi

https://doi.org/10.37546/JALTJJ44.2-5

JALT Journal, Vol. 44, No. 2, November 2022

松岡京一
Koichi Matsuoka
宮城県築館高校
Tsukidate High School, Miyagi

高校の授業内スピーキングテストにおいて、シンプルなルーブリックを用い、詳細な採点者トレーニングを行わない場合に採点者信頼性が十分確保できるかを、グループ型のディスカッションとディベートで検証した。227名の高校生の発話をそれぞれ教員2名で採点し、多相ラッシュ分析・一般化可能性理論等で分析した。その結果、採点者間一致度・一貫性と採点者内一貫性の観点で十分な信頼性が満たされていることが示された。グループ型タスクの場合に、ルーブリックにやり取りの適切さでなく言語面の観点を入れる、生徒が話す時間を長めに設定する、生徒の役割や発言する順番を決める、共通認識がある教員で採点を行う等の信頼性を高める方法とその問題点が示唆された。

Securing rater reliability for classroom speaking tests can be difficult because teacher-raters typically do not have much time to engage in rater training to understand and discuss rubrics and scores. Furthermore, a teacher typically faces difficulties asking colleagues to help double mark each student's performance. Intensive rater training and double scoring are typical procedures to maintain high reliability (Knoch et al., 2021) but are not well practiced in the classroom. However, in some cases, extensive training or double scoring is not necessary when teachers use a rubric with a few criteria and levels, which is simpler than conventional detailed rubrics (Koizumi & Watanabe, 2021). Thus, we use a group discussion and a debate to explore rater reliability when Japanese senior high school teachers use simple analytic rubrics without detailed rater training. We pose the following research questions (RQs):

RQ1: To what degree are raters similar in terms of interrater consensus and consistency?

RQ2: To what degree do raters score students' responses consistently?

RQ3: How many raters are required to maintain reliability?

We analyzed ratings for two speaking tests administrated in September or November to 227 third-year students at a public senior high school. Each test, taken by a group of four students, included either a five-minute group discussion or a 21-minute group debate; the test administration and marking were conducted during the lesson time. An analytic rubric was developed for each task and consisted of three or four criteria with three levels (e.g., content, expression, and technique). Two of the three raters scored each student's response during the test. Teachers had no time to discuss the rubrics in detail and engaged in only a 10-minute discussion about the rubrics before the tests. The ratings were analyzed separately for each test using weighted kappa statistics, Spearman's rank-order correlations, many-facet Rasch measurement (MFRM), and multivariate generalizability theory (mG theory).

The results indicated that the overall rater reliability was adequate, but some cases required careful training. For RQ1, the kappa statistics of two raters' scores for each criterion ranged from poor to substantial agreement (-.06 to .84). Correlations between two raters' scores ranged from negligible to strong (-.07 to .91) and there were not large differences in rater severity (i.e., differences in fair mean-based average values of 0.07 to 0.16 with full marks of 3). In addition, the overall agreement percentages from MFRM were higher than those predicted by MFRM (e.g., 72.9% > 71.6%). The intrarater consistency examined for RQ2 using Infit and Outfit mean squares from MFRM was also adequate (e.g., 0.86 to 1.35). The number of raters needed to maintain sufficient reliability (Φ = .70) for RQ3 was one at the overall test levels and one to three at the criterion levels.

Using simple rubrics, a group discussion task, and a debate task, the results showed that rater reliability can be maintained without extensive rater training. Although the current results may have been affected by study contexts, such as procedures and students' and raters' characteristics, they provide pedagogical and methodological implications for developing speaking assessment tasks and procedures and reporting rater reliability statistics from multiple perspectives.

Keywords: シンプルなルーブリック; 採点者トレーニング; 多相ラッシュ測定; 多変量一般化可能性理論; simple rubrics; rater training; many-facet Rasch measurement; multivariate generalizability theory

スピーキング指導の成果確認のために教室内でスピーキング評価を行う際には、様々な困難が伴う。例えば、実施や採点に時間がかかる。採点時にはルーブリック（採点基準）と生徒の発話を関連づけ、適切なスコアを付けることが求められる。

文部科学省（2020）を整理して再分析した結果によると、2019年度に1年間スピーキングテスト（ST）を実施しなかった（または実施する予定がなかった）科目の割合は、中学校で5.9%、高校では47.8%に上っていた。またSTで使われたタスク形式を見ると、面接とスピーチ、プレゼンテーションがほとんどで、中学校では84.4%、高校では92.1%を占めていた。生徒同士でやり取りを行うタスク（やり取りタスク）として、ディスカッションとディベートは典型的であるが、2つを合わせて中学校で9.6%、高校では4.7%しか使用されていなかった（文部科学省、2020）。

スピーチやプレゼンテーションは発表の力を測り、一般的な面接は、教員と話すことでやり取りを行う力（やり取り力：interactional competence）を測るが、それらの形式だけでは測るやり取り力が限られる。やり取り力は、話し手が他者とやり取りする際に用いる力で、コミュニケーション力の重要な要素である（Celce-Murcia, 2007）。やり取り力には会話を開始する、話題を変える、トピックやターン（発話権）を管理する、やり取りが止まったときに相手を助けつつ会話を何とか続ける（修復する）などの力が含まれる。STに教員と話す形式だけでなく、生徒同士で話す形式を入れることで、学んだ項目や機能を幅広く適切に使う力を測ることができ、また指導時のタスクと近

いものを使うことで動機づけや習得によい影響を与えられるとされる（小泉、2018c; Galaczi & Taylor, 2021）。

2020年度から順次施行されている学習指導要領（文部科学省、2019）において「話すこと」が「やり取り」と「発表」に分かれて記述された。それに伴い、やり取りを意識した指導が積極的に行われていくだろう。しかし、それを支える評価の基盤は弱く、改善が必要である。髙島（2019）が中学校の教員に行ったアンケートによると、STの実施に自信がない教員は、やり取りで57.14%、発表で38.10%おり、発表よりやり取りのテスト実施に自信がない傾向が見られる。さらにST実施後の採点に関して自信がない教員はさらに多く、やり取りで61.90%、発表で66.67%であり、約3分の2を占めている（p. 16; %は筆者らが計算）。STをより多く実施している中学校においてこの結果であるため、高校ではSTの実施や採点に自信がない教員がさらに多い傾向が見られると考えられる。そのため、やり取りの評価、特に実施と採点に関する研究と実践が求められている。本研究では、やり取り力の適切な評価に向けて、やり取りテストの採点に焦点を当て、高校の授業内に行ったディスカッションとディベートのテストを採点する際に、シンプルなルーブリックを用いた場合の採点の質を調べる。

採点者信頼性の種類と対策

教室内で行うSTの採点では、テストで生徒が話した英語を聞き、ルーブリックと突き合わせてどのレベルにあたるかを、教員が採点者（評価者・評定者）として判断するのが一般的である。採点者は、採点間で一致するような判断を安定して行うこと（高い採点者信頼性）が求められるが、意図せずにルーブリックに沿わない判断をすることがある。採点者信頼性は採点者間と採点者内の観点で分けられる（Luoma, 2004）。採点者間信頼性は、異なる採点者の間で同じような厳しさやパターンで採点しているかである。採点者内信頼性は、同じ採点者が採点中同じように採点しているか、また同じスピーキング力を持つ受験者を同じように採点しているか（例：他教科の成績がよい生徒のスコアを、英語の発話に関わらず高くしていないか）、複数の観点を別々に判断する分析的（analytic）採点であれば、採点観点で厳しさを違えて採点していないか（例：文法より流暢さの方が難しくなるように意図したルーブリックで、他の採点者はその方向で採点している中、文法の方が難しくなる方向で採点していないか）等の観点で調べられる。

Stemler（2004）によると、採点者信頼性は3つのアプローチで分類できる。第1の一致度（consensus）アプローチでは採点者間のスコアが一致しているかを調べ、単純に一致した割合を出す一致率や、偶然起きる一致を調整した値であるkappa係数などで示す。第2の一貫性（consistency）アプローチでは、採点者同士が同じ傾向で採点しているかを調べ、Pearsonの積率相関係数やSpearmanの順位相関係数、Cronbachのアルファ係数等で示す。第3の測定（measurement）アプローチでは、主成分分析や多相ラッシュ分析（many-facet [またはmultifaceted] Rasch measurement: MFRM）、一般化可能性理論（generalizability theory: G theory）などの測定モデルを使って調べる。MFRMでは、採点者の厳しさ（rater severity）値や、採点者間の一致率、採点者の採点パターンがラッシュモデルから予測されるパターンと一致しているかを示す「採点者適合度指標値」等が算出される。この中で、採点者の厳しさ値と一致率は、概念としては第1の一致度アプローチに近い。採点者適合度は第2の一貫性アプローチに

近く、採点者内一貫性を示す。表1では、採点者2名または採点2回の場合に使える採点者信頼性の指標を、本研究で使用するものを中心に整理した。目的によるが、採点者信頼性を検討する際には、より包括的に検討することが望ましいとされる。

教室内STでは、大規模テストや重要な判断に用いるテストとは違い、非常に高い信頼性は求めなくてもよい。しかしある程度の信頼性は保たれるべきで、系統的に信頼性が低くならないような方策をとる必要がある（Knoch et al., 2021; Luoma, 2004）。その典型的なものとしては、以下3つの方法がある。第1に、採点前に採点者トレーニング（またはstandardization, calibration, moderation）として、採点者がルーブリックやサンプル発話、スコア、その理由等を確認した後、別の発話を聞いて採点者が個々に採点し、その後ずれをなくすために話し合い、調整する方法である。第2に、本番のテストの発話を2名以上で採点し、その後ずれがある場合に話し合う、平均点を使う、別の採点者が採点する等のプロセスを経て、最終スコアを決める方法である。第3に、1名の採点者が時間をおいて2回行う方法である。

上述の採点者信頼性を担保するための3つの方法は、教室内STでは実施ができないことが多い。採点者トレーニングは最低でも1時間、徹底的に行う場合には数時間かかるが、その時間は確保できないことが多い。複数の採点者の確保は、外国語指導助手やティーム・ティーチング配置がない場合には難しいことが多く、同じ採点者が2回最低するのも採点に2倍の時間がかかることになり、同様である。

信頼性確保のための一般的な手順が満たせない場合に、その手順を行うための環境づくりに注力する方向もあるが、他の方法がないかを考える方向もある。本研究は後者のアプローチをとり、ルーブリックの簡素化を試み、その影響を探る。

表1.
採点者2名または採点2回の場合の採点者信頼性指標

観点	主な指標	解説	基準・解釈
採点者間一致度	一致率	一致したスコアの割合	0〜100%で、100に近いほど一致度が高い
	kappa係数（κ）	偶然起きる一致率を調整した値。値は-1〜1。基準の英語は、低い順から、none to slight, fair, moderate, substantial, almost perfect。0以下は一致なし（no agreement）	.01〜 .20:若干 .21〜 .40:まずまず .41〜 .60:中程度 .61〜 .80:十分 .81〜1.00:ほぼ一致[a]
	一致率【MFRM】	一致したスコアの割合。予測一致率は採点者の厳しさ値を考慮して算出	予測一致率よりも少し高いのがよい[b]
	採点者の厳しさ（推定）値の差の最大値【MFRM】	厳しさ値の最も離れた採点者2名の値の差。どの程度の違いがあるかを示す。厳しさ値は、受験者やタスク等の影響を調整した値。ロジット尺度[c]上で表され、0が平均値、プラスの値は採点者が厳しいことを示す	絶対値で0に近いほどよい。他の相の差よりかなり小さいと差が小さいと考える
	採点者（分離）信頼性【MFRM】	採点者の厳しさ値にどの程度の違いがあるかを示す	0〜1の値を取り、高いほど異なる。0に近いほどよい
	採点者のfair scoreの差の最大値【MFRM】	fair score（fair mean-based averageの値）は、採点者の厳しさ値を素点の尺度に直した値。最も離れた採点者2名の値の差。実質的にどの程度の差があったかの判断材料になる	明確な基準はないが、差が大きければ、素点を使う場合には実質的な影響が出ると解釈する
採点者間一貫性	Spearman順位相関係数（r_s）	2名の採点者の採点パターンが似ているか（例:よい発話をした受験者に、採点者がともに高いスコアを付けているか）を示す。値は-1〜1	.25〜 .39:弱 .40〜 .59:中程度 .60〜1.00:強[d]
採点者内一貫性	採点者適合度指標値【MFRM】	採点パターンが、ラッシュモデルから予測されるパターンと一致しているか。Infit平方平均（mean squares: MS）とOutfit MSがあり、通常は前者で判断。値は0〜無限大	例:〜0.49:過剰適合 0.50〜1.50:適合 1.51〜:不適合（2.01〜は測定に影響する可能性がある）[e]
その他の採点者信頼性	採点者分散の割合【G theory G研究】[f]	スコア全体の分散の中で採点者分散が占める割合。採点者がスコアに与える影響の度合いを示す	小さいほどよい
	必要な信頼性を満たす採点者数【G theory D研究】	テストの信頼性（G係数・Φ係数）が、基準値[g]以上になる場合で判断	小さいほど安定して測定できている

注. 包括的な指標はKnoch et al.（2021）、McKay and Plonsky（2021）を参照。A〜B = A以上B以下。【 】= 分析法。[a] McHugh（2012）。[b] Linacre（2021）。[c] 受験者がタスクに成功する確率を基に算出した尺度で、ロジット（logit）を単位とする。[d] Plonsky & Oswald（2014）。他の基準も存在する（例:嶋田、2017）。[e] 基準の英語は、低い順からoverfit、fit、underfit（またはmisfit）。overfitとunderfitを合わ

せてmisfitと言うこともある。目的に応じて基準を変えることもできる（Wright & Linacre, 1994）。[f] 複雑なデザインの場合には、採点者相と他の相との交互作用の割合も関わる。[g] 基準値はテストの重要性を考慮して決める。例えば教室内テストは .70以上、重要性が低い標準化テストは .80以上または .85以上、重要性が高い標準化テストは .90以上とされる（Wells & Wollack, 2003）。

採点者信頼性の現状

Jönsson et al.（2021）は、教師間の成績の付け方にはばらつきが大きい傾向があることを、先行研究に基づき述べている。また彼らは実証研究を通して、卒業という重要な判断が成績のみで行われるスウェーデンにおいて、外国語としての英語の成績で教師間の信頼性は高くないことを示した（例：スコアの中央値との一致率：100%が望ましいところで59.7〜66.7%）。

日本においては、信頼性を研究トピックとすることや、信頼性を量的研究の一部として報告することは限られている。例えばStapleton and Collett (2010) は*JALT Journal*の過去30年間の論文の中で、テストの信頼性と妥当性を扱った論文は少なく（5.72%, 17/297）、量的研究の中で信頼性を報告した研究も少ないことを明らかにしている（10.53%, 8/76）。McKay and Plonsky (2021) によると量的研究の中で信頼性を報告する研究の少なさの傾向は国際誌でも同様である（例：16〜40%）。

このように研究としての信頼性の報告が限られる中で、日本の実際の教育活動の中で採点者信頼性を調べた実践はさらに限られるだろう。研究的な側面はあるが、中高生とその教員対象に行った研究が数件ある。例えばAso (2000) では、英語教員10名が高校生10名の英語面接時の発話を、分析的・総合的（holistic）ルーブリックを用いて採点した（採点者トレーニングやルーブリックの詳細提示の記述はなし）。その結果、2つのルーブリック両方で、採点者間の相関が低いものから高いものまであり、採点者間一貫性は一部の採点者の間でのみ満たされていた（例：総合的でr_s = .26〜 .96）。一方、同じ採点者が半年間を空けて2回行った採点を比較したところ、採点者内一貫性は非常に高かった（例：総合的でr_s = .98）。

採点者信頼性をMFRMで調べた研究もある。大学生を含む日本在住の英語学習者を対象にした研究を表2にまとめた。例えばNegishi (2011) では、日本人中学生から大学生までの135名がグループ型ディスカッションテストを受け、11名の日本人高校・大学教員が3日間程度の採点者トレーニングを受けた後に135名の発話を採点した。MFRM結果では、採点者の厳しさ値に違いがあり（厳しさ値の差：ロジット値で3.25）、適合度（Infit MS）では1名が0.50〜1.50の範囲内に入らなかった（1.94）。その1名を除いた再分析でも別な採点者2名が若干の問題を示した（Infit MS = 1.51, 1.55）。なお、表2のまとめにおける採点者数はテスト採点に関わった人数であり、一般的には1人（1組）の発話は2名で採点されていることに注意したい。例えばVan Moeren (2006) では1グループ4名の会話を採点者40名中の2名が聞いて採点した。

表2.
MFRMを用いた採点者信頼性研究（日本の英語学習者対象に限る）

	受験者	タスク形式	採点者	分析的ルーブリック観点数	トレーニング時間	厳しさ値の差[a]	採点者適合度[b]
Sato (2012)	大学生156名	意見表明	9名	5個5段階[c]	なし[d]	あり[e]	範囲内
Inoue (2013)	大学・院生65名	絵描写	9名	5個5段階[c]	3時間	あり/なし (1.78～2.11)	範囲内
Hirai & Koizumi (2013)	大学・院生48名	技能統合型再話	9名	3個5段階	1～2時間	あり[f]	範囲外1名
Yokouchi (2018)	大学・院生128名	技能統合型再話	4名	4個5段階[g]	20分と個人練習[h]	なし (0.17～0.21)	範囲内
Akiyama (2001)	中学生109名	面接[ij]	4名	5個5段階	約1時間	なし (0.72)	範囲外1名
Akiyama (2004)	中学生288名	面接[k]	10名	5個6段階	2時間	あり (3.84)	範囲内
Iwamoto (2018)	大学生46名	面接[l]	4名	4個9段階	不明	あり[l]	範囲内
Nitta & Nakatsuhara (2014)	大学生30名	ペア型[m]	2名	3個9段階	90分	なし (0.0)	範囲内
松村・守屋 (2019)	大学生38名	ペア型[m]	2名	4個5段階	8時間[n]	なし(0.32)	範囲内
Koizumi et al. (2020)	大学生110名	ペア型[km]	3～4名	4個3段階[c]	5～8時間	なし(1.18～1.42)	範囲内
Nakatsuhara (2007)	高校生42名	グループ型[m]	2名	5個6段階	1時間	なし (0.06)	範囲内
McDonald (2018)	大学生64名	グループ型[m]	4名	5個9段階[o]	2時間	なし (1.27)	範囲内
Bonk & Ockey (2003)	大学生1103～1324名	グループ型[m]	20～26名	5観点9段階	2時間	あり (最大で4.50)	範囲外約4～7名
Van Moere (2006)	大学生113名	グループ型[m]	40名	5観点9段階	90分	あり (3.41)	範囲外6名
Negishi (2011)	中学～大学生135名	グループ型[m]	11名	5個7段階[c]	3日間程度	あり (3.25)	範囲外1名
Negishi (2015)	大学生24名	ペア・グループ型[m]	5名	総合的10段階	3日間程度	なし (0.62)	範囲内

注．企業作成のテストの研究と採点者相分析がない研究は除く。[a] 最大値がロジット尺度で2以上の場合を差ありとした。[b] 値掲載がある場合にはInfit/Outfit MSの0.50～1.50の間を範囲内とした。掲載がない場合には、論文の記述に沿った。[c] 総合的ルーブリックも使用。[d] 採点手順やサンプルでの練習資料は提供。[e] 採点者分離信頼性 = .99。[f] 採点者Separation = 2.32。[g] 4個5段階に0を加えた計21段階で分析。[h]（横内、私信、2021年3月8日）。[i] 絵描写。[j] ロールプレイ。[k] ペア型ロールプレイとスピーチ。[l] 採点者Separation = 4.04。[m] ディスカッション。[n]（松村、私信、2020年10月5日）。[o] 5個5段階の結果も同様。

表2から、中高生対象のSTの研究が少ないこと、分析的ルーブリック観点は3～5個、段階はKoizumi et al.（2020）以外は5～9個と多いこと、採点者トレーニングは行う場合は1時間以上が多いことなどが見えてくる。厳しさ値については、差がある場合とない場合があり、採点者適合度はどのテスト形式でも満たす採点者が多い。しかし、発表型の技能統合型再話やグループ型までどの形式でも、適合しない採点者はトレーニング後でも見られる。先行研究では、トレーニングや個別フィードバックを行って採点者の一致度や一貫性が改善した例とそうでない例があり（McNamara et al., 2019）、教室内テストに限らず採点者が関わるテストでは課題となっている。

表2の採点者適合度では、特にグループ型のBonk and Ockey（2003）とVan Moere（2006）での範囲外の採点者の多さが目を引く。これは受験者数や採点者数が多いためもあるだろうが、グループ型の採点が難しい可能性もある。後で詳細に述べるKoizumi and Watanabe（2021:以後K&W）では、採点者トレーニングがほぼない場合にグループ型採点の難しさを支持する結果が出ている。グループ型では一般に、3名以上の受験者がいつ話すか分からない状況で採点するため、1～2名のときよりも採点者の認知的負担が高く、難しい可能性がある。

McNamara et al.（2019）によると、MFRMを用いれば採点者の厳しさ値の違いを調整したスコアが出せる。しかし、不適合の採点者の影響はMFRMでも調整できず、問題となる。また教室内テストでは、採点スコア（素点）をそのまま使うことが多いため、採点者の厳しさ値の違いも検討事項となる。

表2の中でG theoryも行った研究において、十分な信頼性を保つために最低必要な採点者数は、タスク2個で1名（松村・守屋、2019）、3観点で2名（Akiyama, 2001）、テスト1回で4名（Van Moere, 2006）と様々だったが、8時間のトレーニングを行った松村・守屋（2019）を除くと、通常2名は必要だった。

まとめると表2で挙げた研究では、観点は3～5個で、5段階以上が多いなど詳細なルーブリックを使い、採点者トレーニングは1時間以上行うことが多い。しかし、教室内評価の場合、生徒の能力の幅は狭く、指導目標を達成したかの確認が重要なため、焦点を絞った少ない段階のシンプルなルーブリックを使うので十分という考え方もあるだろう。その場合、測れる力が限定されたり、結果の診断機能が少なくなって指導や学習に使える情報が減ったりという問題もある。一方、シンプルなルーブリックを使うことで採点者トレーニングを詳細に実施しないとしても、また採点者を2名確保できないとしても、十分な信頼性が保たれたり、採点の負担が減って実行可能性が高まったりするならば、年間で数回定期的に行い、採点者信頼性をある程度保ちたい状況では、この方が教室内STに適しているという考え方もある。

この考えに基づいて行ったK&Wでは、3観点、3段階のシンプルな分析的ルーブリックを使用し、採点者の事前の打ち合わせを10分間のみ行い、テストの最初の2～3名

（2〜3組）を独立に評価した後にずれや疑問点を話し合い、基準を調整した。その上で採点を複数名で行い、どの程度採点者信頼性が保てるかを調べた。授業に即したテストを年4回行い、採点者は2〜9名で授業時間のテスト中に採点を行った（表3参照）。採点者の厳しさは、fair scoreの差の最大値で見ると、第2回グループ型ディスカッション以外はテストの3点満点中0.50未満で実質影響がない範囲と考えられた。採点者適合度は、ペア型ロールプレイ以外では問題が見られず、一致率は予測一致率より高い結果で、全体的には採点者一致度と一貫性がほぼ満たされていた。しかし観点ごとの一致度を見ると、個人プレゼンテーションとペア型ロールプレイはまずまずだったが、2回のグループ型ディスカッションでは低かった。G theoryの結果では、1〜4名が十分な信頼性を保つために必要という結果になった。全体的には、採点者トレーニングがない割にあまり問題がなく、十分運用ができる範囲で、一部、特にグループ型ディスカッションで注意が必要と考えられた。

表3. Koizumi and Watanabe（2021: K&W）のテスト内容と採点者信頼性結果

タスク形式	個人プレゼンテーション	第1回グループ型ディスカッション	ペア型ロールプレイ	第2回グループ型ディスカッション
実施時期	7月	10月	12月	1月
測る力	発表＋やり取り	（発表＋）やり取り	やり取り	（発表＋）やり取り
使用授業回数	1	2	2	1
採点者の厳しさ値の差	0.82	1.65	1.47	3.35
fair scoreの差の最大値[a]	0.22	0.29	0.43	0.98
採点者適合度	範囲内	範囲内	ほぼ範囲内[b]	範囲内
一致率（MFRM）	予測一致率より高かった			
観点ごとの一致率[c]	61.0〜75.2%	50.9〜72.7%	77.1〜81.7%	47.8〜57.5%
観点ごとのkappa係数c	.45〜 .71	.17〜 .40	.79〜 .82	.10〜 .54
Spearman相関[c]	.45〜 .74	.18〜 .44	.76〜 .82	.11〜 .59
合計点の信頼性確保に必要な採点者人数	2	4	1	3

注. [a] 3点満点中。[b] Outfit MSを若干外れた採点者が2名（0.49と1.88）。[c] 本研究のために計算（G theoryで使用したデータを使用）。

この結果は、他のタスク形式ではどうなるだろうか。本研究では、K&Wでも用いたグループ型ディスカッションとともに、やり取り力を測るタスク形式としてグループ型のディベートを用いる。本研究により、K&Wと比較しながら浮かび上がる、多様なタ

スク形式での採点者信頼性を維持し、教育の資源を適切に配分するための手順が明確化されると思われる。

目的と研究課題

本研究の目的は、高校生のグループ型のディスカッションとディベートを採点する際に、詳細な採点者トレーニングがなくシンプルなルーブリックを用いた時の採点者信頼性を調べることである。研究課題は以下3点である。

研究課題1:採点者間一致度・一貫性の点で、採点者はどのような採点を行っているか?
研究課題2:採点者内一貫性の点で、採点者はどのような採点を行っているか?
研究課題3:十分なテスト信頼性を持つために、何人の採点者が必要か?

方法

受験者と採点者

受験者は、日本の公立高校の3年生227名であり、STは必修の英語の授業の中で受験した。対象校は地域の進学拠点校であり、受験者は6クラスのうちの1クラスに所属していた。このクラスの授業では、コミュニケーション力を高めることを目的として4技能を用いる活動が普段から多く行われていた。受験者の4技能の英語力は、CEFR-J(日本版ヨーロッパ言語共通参照枠;投野・根岸、2020)のA2.1が35%、A2.2が47%、B1.1以上が11%であり、スピーキング力は、A2.1が37%, A2.2が36%, B1以上が0%だった(7月のGTECの4技能テスト結果[3技能版で219名、STで224名受験]に基づく。Benesse Corporation, 2019参照)。

STの採点者3名は全員、同じ学年で同じ科目の授業を分担して担当していた教員である。採点者3名のうち、クラスごとに異なる2名がペアとなって採点を行った(授業担当者ともう1名。当日の事情により、生徒2名のみ教員1名で採点)。3名の教員は日本人で、10年以上の英語指導歴があった。3名は、定期的に「CAN-DOリスト」の形での学習到達目標を確認し、授業前に指導理念や方法、教材を共有していた。一方、テストタスクとルーブリックについては、事前に10分ほど話し合いを行ったが、サンプル発話の採点や詳細な議論などの精密な採点者トレーニングは行わなかった。

テストタスクとルーブリック

生徒は年間で、コミュニケーション英語IIIで3回、英語表現IIで3回、計6回のSTを受けた。その中で、英語表現IIの2回目(9月)と、コミュニケーション英語IIIの3回目(11月)のテストが今回の分析対象である。対象生徒が受けたSTの中で、外部研究者である第1著者に情報開示が可能だった採点データを分析した。

テストタスクやルーブリックは、第1著者のアドバイスのもと、授業担当者3名のうち1名が作成した。タスク形式やルーブリック、またトピックについてはすべて、事前に生徒に提示した(表4、付表A・B・C・D参照)。教員は、テストに向けてどのように準備するかは生徒に詳細には示さず、授業で学んだ表現をテストで使えるように復習してお

くように伝えた。STは生徒全員が教室にいる状態で行われ、実施や採点はクラスごとに2名の教員で行った。

一般に言語テスティング研究におけるタスクは、発話を引き出すために受験者に提示される活動を意味するが、本研究におけるタスクは第二言語習得研究で習得に役立つものとして挙げられている4点（Ellis & Shintani, 2014; 福田他、2017）、活動中の焦点が意味にあること、発話者間に情報のギャップがあること、発話者のリソースを使ってタスクが実行されること、タスク達成が（内容の観点で）評価されることを満たしていた。

表4.
スピーキングテスト（ST）の詳細

タスク	ディスカッション（9月）	ディベート（11月）
科目	英語表現II	コミュニケーション英語III
測る力	やり取り	やり取り
グループ形式	4名で1グループa。司会者1名と参加者3名	4名で1グループ[a]。4名それぞれに立場（役割）を割り当てあり
時間	1グループ約5分 50分授業の1.5回分を使用	1ディベート2グループ参加、21分（話す時間は11分） 50分授業の2.5回分を使用
分析的ルーブリックの観点（3段階、計30点[b]）	内容（10点） 表現（10点） 技術点（10点） 内容と表現について、司会者と参加者で異なるルーブリックを使用	内容（12点） コミュニケーションに対する姿勢（8点） 文法・語法（5点） 音量・速度・発音（5点） 内容について立場ごとに異なるルーブリックを使用
トピック例[c]	1. Why do you think some students study abroad? 2. Which do you prefer to work for a large company or a small company?	1. Students should be asked to study foreign languages other than English. 2. There should be boys-only and girls-only high schools in addition to co-education schools.

注. [a] 60グループ中、数グループは欠席等のために3名または5名で構成された。[b] 評定100点中の30点は、それぞれのST結果を使って決定された。[c]トピックはそれぞれ11個と12個で事前提示。付表C・D参照。

ルーブリックは、3〜4観点の3段階（レベル1〜3）の形で作成した（表4・付表AとB参照）。ディスカッションとディベートの観点は、「内容」は共通で、「技術点」は「文法・語法」と同じ、「表現」は「コミュニケーションに対する姿勢」と「音量・速度・発音」に分岐という形で、ラベルが異なっていても共通の観点で構成されていた。

グループ型ディスカッション

このタスク形式では、生徒はトピックに基づき、グループごとに約5分間英語で話し合った。トピックごとに司会者1名と参加者3名と、各参加者が発言する順番を予め決めておき、1人目→2人目→3人目→1人目の順で発言することとした。司会者は司会進行時に自分の意見を述べても構わないとした。参加者は、直前の参加者の意見に言及してから自分の意見を述べることとし、1人目の参加者も2回目の機会にそれを行うこととした。トピック以外の指示はなく、5分以内に合意に達するよう求めることもなかった。

テストの授業前には、授業担当者が事前に作成した1グループ4名、1クラス10グループのグループ分けを全員に向けて発表した。その後、グループ内で司会者1名を決めた。司会者はトピックごとに交代し、テスト全体で1人あたり計2〜3回担当した。

テスト中にトピックを提示した後に、話す内容や表現を考える時間はなかった。教員が1つのトピックを提示すると生徒はすぐにディスカッションを始め、5分が過ぎた時点で止め、再び教員が次のトピックを提示すると、同じグループで役割を変えて次のディスカッションを開始する、という形式で進めた。各グループが（事前に提示していた11個のトピック中の）10個のトピックについて話し続ける中、教員2名が1個のトピックについて1つのグループを観察し、次のトピックでは時計回りに次のグループへ移動して評価を続けた。教員の採点を始める位置は異なっており、教員は毎回異なるトピックの会話を採点した。生徒にとっては、あるトピックでディスカッションを行っている時に教員1名が近くで採点しており、数個後の異なるトピック時に別な教員1名が近くで採点している形だった。10回のディスカッションのうち2回が採点対象だったことになる。

本手順は、松尾（2019）を若干修正したものである。本研究の手順の利点は、生徒が授業中に継続してディスカッションを行うことになり、スピーキングの機会が十分確保できること、教員にとっては、生徒の役割と発言する順番が分かっていることで、今話している生徒の発話に集中できることである。弱点は、採点の対象になるトピックや順番（例：1回目のディスカッションでの採点と、慣れてきた時点での採点）がグループごとに異なることである。このことは採点者信頼性の観点からみると、教員の採点が一致しなかった際には、教員間の採点のずれからの影響だけではなく、採点のタイミングやトピックからの影響も考慮する必要があるということを意味する。しかしトピックについては、事前提示があり準備ができる形になっていたため、また似たタスクの先行研究ではトピックのスコアへの影響は見られなかったため（例：Van Moere, 2006）、影響は少ないと思われた。

グループ型ディベート

このタスク形式では、表5の流れに沿って、1グループ4名から成る2グループの試合形式で、準備時間を含めて21分間ディベートを行った。どちらのグループが勝利したかについては、ディベートをオーディエンスとして聴いたクラスの生徒の挙手数によって決定した。2グループごとにディベートを行い、それを異なるトピックで5回繰り返した。

表5.
ディベートの流れ

	Stage	Team A	Team B
準備4分	① 肯定側立論(90秒)	論題を肯定する立場でメリットを述べる	
	② 否定側立論(90秒)		論題を否定する立場でメリットを述べる
準備2分	③ 質疑(60秒)	相手側チームの意見に質問をする	
	④ 質疑(60秒)		相手側チームの意見に質問をする
準備2分	⑤ 反論(90秒)	立論・質疑を踏まえたうえで、再度自分たちの優位性を説明する	
	⑥ 反論(90秒)		立論・質疑を踏まえたうえで、再度自分たちの優位性を説明する
準備2分	⑦ 総括(90秒)	自分たちの主張の方が重要性が大きいことを印象付ける	
	⑧ 総括(90秒)		自分たちの主張の方が重要性が大きいことを印象付ける

注. 準備時間中は、グループ内で話し合いが行われた。

テストの授業前には、授業担当者が事前に作成した1グループ4名、1クラス10グループのグループ分けを全員に向けて発表した。試合を行う2グループ、トピック、グループごとの肯定側・否定側の立場は、ディベート開始直前に授業担当者がくじをひいて決定した。グループ内での立場の割り当てはグループ内で決めた。表5の流れに沿ってディベートが行われ、教員はそれを聞きながら採点を行った。

本手順は、教科書のディベート活動に基づくもので、その長所は、生徒は他のグループの様子を確認することで英語表現やディベートの効果的な方法を体感することができる点である。また教員にとっては、ディスカッションと同様に、生徒の役割や発言する順番が決まっていることで、発話中の生徒に意識を向けやすく、採点に集中しやすい。また2名の教員が同じ発話を採点するため、スコアがずれた時の理由の特定や、スコアの調整が行いやすい。一方短所は、後に行うグループは、事前提示のトピックのどれが提示されるかは分からなくても、他グループのテスト中にある程度は準備が可能で、また他グループの様子からも学ぶことができ、有利になりやすい点である。

タスクと授業目標・指導の関係

2つのタスクとも、授業目標と事前の指導で用いたタスクに基づいて作成した。STでのトピックは指導時に用いたタスクに近く、生徒の興味をひくもので、生徒の現在または今後の生活に関係するものという視点で設定した。

指導やテストで使用するタスク形式やトピックを決める際には、この学校が定める、高校3年生のやり取りの領域における「CAN-DOリスト」の形での学習到達目標も参照した。高校3年生後期の目標は「社会問題や抽象的な話題について、流暢かつ自然に対話ができると共に、建設的な議論の構築に積極的に参加し、相手を説得できるように自分の考えを説明することができる」であった。高校3年生前期のやり取りの学習到達目標は「社会問題や抽象的な話題について，相手を説得できるように説明したり，情報を交換したりすることができる」であり、前期と後期の目標は、後期目標中に下線を引いた部分が大きく違っていた。前期実施のSTと指導時の反応に基づくと、前期終了の時点で前期の目標は8割の生徒がおおむね満たしたと思われた。後期には、後期の目標を達成するために、自分の意見を述べさせる活動を多く行った（まずペアやグループで話し、次に書く形式）。その際、(a)　自分の考えや意見を根拠・理由、具体例を添えて話すこと、(b)　相手に伝わるように発音や文法、語法、またアイコンタクトなどを工夫して表現することを強調した。これらのポイントは、分析的ルーブリックの (a) 内容と (b) それ以外（例:ディスカッションの観点では「表現」と「技術点」）に反映させた。

1グループ4名でのディスカッションも、1グループ4名、2グループでのディベートも、テスト前の授業中に実施し、生徒は形式に慣れていた。特に2つのテストの前の授業では、通常の授業の延長として役割（立場）も含めてテスト本番を想定した練習を複数回行った。その際にはテストのトピックとは別のものを用いた。

採点

上述のように、3名の教員間で事前に詳細に話し合う採点者トレーニングを行う時間はなく、10分程度の情報の共有のみを行った。1名の生徒につき、2名の教員が独立に採点した。テストの録画・録音はされなかった。

採点時には、すべて生徒一人ひとり観点ごとに採点したが、ディベートの「内容」は個人点を付けた後、4人グループでの平均値を出し、内容グループ点も算出した。内容についてはグループでの準備時間の話し合いが反映されており、グループ全体の力が反映されていると考えたためであった。本研究の分析では、個人点とグループ点を使った場合の両方を分析した。成績には、2名の教員のスコアの平均値を使用し、ディベートの内容はグループ点を用いた。結果はスコアレポートとして生徒に返却した。

なお、内容グループ点を成績点に含めることが適切かは議論を要する点である。生徒一人ひとりの発話とスコアに大きく影響するのが、個人の力とグループの力のどちらと考えるかによって捉え方が異なってくる。言語テスティング研究において、生徒のやり取りでの発話やスコアが何を意味するのか、発話やスコアを個人の力として捉えてよいのか、対話者や採点者等の要因の影響をどの程度受けるのかについては長年議論がされている（Iwashita et al., 2021; McNamara, 1997）。今回は、発話直前に提供された準備時間では、グループで内容は相談できるが表現を話し合う時間はあ

まりなく、内容のみグループの影響が大きいと考え、内容グループ点を成績に採用した。しかし、それを用いることでの影響を調べるため、内容個人点を使った場合と比較することにした。

分析

もともとのルーブリックには、「表現」のレベルの4点、7点、10点のように各観点に重みづけがあった。またディスカッションの観点はすべて10点満点だったが、ディベートの観点の満点は12点、8点、5点、5点と異なり、その意味でも重みづけがあったが、分析時にはすべて1〜3に変換した。採点者ごと、観点ごとの変換後のスコアを用いて、様々な採点者信頼性指標（研究課題1向け）、MFRM（研究課題1と2向け）、G theory（研究課題3向け）を使って分析した（分析用シンタックスは付表E・F）。重みづけなしの値の1〜3に変換したのは、本研究は信頼性が焦点であり、Linacre（2021）によると、測定値の真の信頼性は重みづけなしの分析から得られ、重みづけを行うことで信頼性の分析に恣意的な要素が入ってしまう（p. 375）ためだった[1]。また重みづけなしに行うことで、ディスカッションとディベートの比較が容易に行えることも理由の1つであった。

様々な採点者信頼性指標はMizumoto（2021）で算出した（Plonsky & Mizumoto, 2021も参照）。採点者の組み合わせごとに（採点者AとB、BとC、AとCの場合で別々に）調べ、採点者間一致度を見るために一致率と重みづけkappa係数を、採点者間一貫性を見るためにSpearmanの順位相関係数を用いた（表1参照）。

多相ラッシュ分析（MFRM; 小泉、2018b; 平井他、2018; McNamara et al., 2019）ではFacets（Ver. 3.83.6; Linacre, 2021）を3回用いた。各観点が個別に機能すると考え、部分採点モデル（partial credit model）を使った。入れた相は、受験者、採点者、ルーブリック観点だった。モデル適合の基準は、テスト単体で重要な判断を行わないため、Infit MSとOutfit MSの0.50〜1.50とした。

G theoryは、mGENOVA（University of Iowa, 出版年不明）を用いて3回分析した。多変量一般化可能性理論（multivariate generalizability theory: mG theory; 小泉、2018a; Grabowski & Lin, 2019）の完全なクロス式で欠損値のない、1相の$p^{\bullet} \times r^{\bullet}$デザインを用いた。採点者をランダム相とし、3〜4個の観点を従属変数として扱った。本分析で採点者は、採点者3名のスコアを圧縮して採点1・2として分析する、G theoryでは一般的な方法（Lin, 2017）を用いた。トピックはグループによって異なっていたが、同じとみなして分析した。G研究で、スコアの分散を、受験者能力の違いに由来する分散、採点者の厳しさの違いから来る分散、その他誤差から来る分散に分けてその割合を出した。D研究では、採点者の数を変化させたときにどの程度信頼性（信頼度：dependability）が変化するかを調べた。本テストは目標基準準拠（criterion-referenced）評価に使われるため、Φ（ファイ）係数を用い、Φ = .70以上を十分な信頼性と考えた。

結果

様々な採点者信頼性指標

表6を見ると、全体としては採点者間の一致度と一貫性が保たれていたが、一部満たされていないものもあった。例えばディスカッションの内容の観点において、採点者AとBで採点が完全に一致したのは74.4%で、kappa係数は .71で十分高く、相関は .70で強い相関があったが、採点者AとCでの完全一致度は57.1%で、kappa係数は .11で若干の一致で、相関は .11でほとんど関係がなかった。

表6.
採点者間信頼性係数:採点者の組み合わせごとの値の範囲

タスク形式	観点	一致率	kappa係数	Spearman相関
ディスカッション	内容	57.1〜74.4%	.11〜 .71	.11〜 .70
(*n* = 21〜117)	表現	52.4〜81.0%	.31〜 .70	.30〜 .73
	技術点	56.0〜81.2%	.06〜 .73	.13〜 .73
ディベート	内容	78.3〜90.6%	.58〜 .84	.61〜 .91
(内容個人点。	姿勢	65.2〜90.6%	.49〜 .82	.54〜 .85
n = 64〜91)	文法・語法	75.4〜90.6%	.40〜 .83	.48〜 .84
	音量等	55.1〜91.2%	.03〜 .82	.06〜 .82
(内容グループ点)	内容	68.8〜86.8%	-.06〜 .56	-.07〜 .63

注. *n* = 採点者組み合わせごとの生徒数。内容個人点 = 内容も他の観点も個人点を用いた場合。内容グループ点 = 内容はグループ点を用い、それ以外は個人点を用いた場合。

採点者2名の組み合わせは24ケースあり、その中の一致度や一貫性が低めだった6ケース(ともに.39未満のものを選択)について、どのようにずれていたかを詳細に調べた(表7参照)。クロス表を見ると、例えばケース1では、ディスカッションの内容の観点においてkappa係数が .11で一致度が低く、相関係数が .11と低かったが、ずれは採点者Aが2と採点したものを採点者Cは3とした場合(4名分)、またはその逆(5名分)であり、レベル2と3でのずれで、ケース4と5と同じパターンだった。ケース2ではレベル1と2、レベル2と3でのずれがあり、ケース3と6ではレベル1と3、2と3のずれだった。この中でより深刻な不一致は、第1に2レベル異なるレベル1と3のずれ(計3件)、第2に授業目標を達成したかを示すレベル1と2のずれ(計2件)であり、採点者トレーニング時や事後に優先して話し合うべき事項だと思われた。また、このように採点者間の採点のずれが大きく現れた場合の発話にどのような特徴があり、採点者がそれをどのように捉えたかについては探求すべき重要な点である。今回はテストの録画・録音がないために確認はできなかったが、それがあれば検討できるだろう。なお、深刻な不一致と考えられるのは、採点全体の中では一部にとどまり、表7の中では5件のみであった。

表7.
kappa係数とSpearman相関係数が低かった場合

	ケース1	ケース2	ケース3	ケース4	ケース5	ケース6
タスク	ディスカッション			ディベート		
観点	内容	表現	技術点	内容グループ点		音量等
クロス表						
一致率	57.1%	52.4%	66.7%	68.8%	86.8%	55.1%
Kappa	.11	.31	.06	.38	-.06	.03
相関	.11	.30	.13	.38	-.07	.06

多相ラッシュ分析（MFRM）

採点者信頼性に関する結果を提示する前に、MFRMの前提であるモデルの全体的適合度の結果を提示する（表8参照）。予想外の回答（Unexpected responses）における標準化残差を用い、±2を超えた標準化残差が約5%以内、±3を超えた標準化残差が約1%以内であれば、データがラッシュモデルに全体的に適合したと考えた（Linacre, 2021, p. 178）。例えばディスカッションでは、それぞれ4.35%, 0.81%で、全体としてこの基準は満たされていた。

表8.
テストの全体的モデル適合度：標準化残差の割合

	ディスカッション	ディベート（内容個人点）	ディベート（内容グループ点）
データポイント数	1,356	1,804	1,804
±2 and ±3を超えた割合	4.35%, 0.81%	3.77%, 0.67%	4.05%, 0.89%

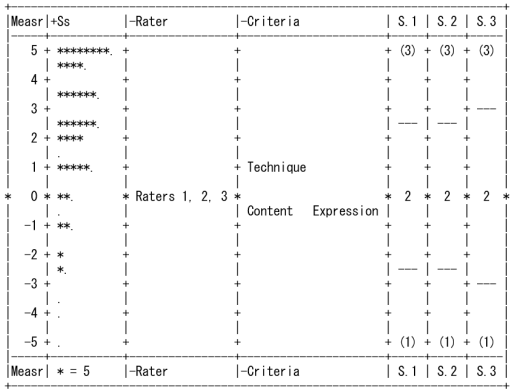
図1の変数マップ（Wrightマップ）の各図において、第1列のMeasrはロジット尺度を、第2列のSsは受験者の能力（推定）値を、第3列のRaterは採点者の厳しさ（推定）値を、第4列のCriteriaはルーブリックの観点の難易度の（推定）値を、第5列以降のS.1～S.4はルーブリックの観点の1～4個目のレベルの分かれ目を示している（値は表9）。変数マップでは通常、値がプラスの方向に高ければ高いほど、受験者の能力は高く、採点者の採点は厳しく、観点の難易度は高く、各観点で高いレベルのスコアを得ることが難しいことを示す。表9の測定値の平均値を比較すると、受験者能力が採点者の厳しさと観点難易度よりも高く（ディスカッションでの例：2.48 > 0.00 = 0.00）、平均的な受験者は、採点者の採点はより甘く、ルーブリック観点はより易しく感じられたと思われる。測定値の標準偏差（SD）では、受験者が最も大きく（2.64）、採点者は最も小

さく(0.14)、採点者の厳しさ値のばらつきは相対的に小さいものだった。標準誤差の平均値では、受験者が最も大きく、採点者と観点はほぼ同じであった。これは、受験者の推定は小さいデータから、採点者と観点の推定はより大きいデータから行うために避けられないことである(概算例:受験者の値の計算は採点者・観点ごとの6〜8個のデータ [2 x 3または2 x 4] から行い、採点者の値の計算は受験者・観点ごとの681〜908個のデータ [227 x 3または227 x 4] から行った。詳細はEngelhard, 2013を参照)。

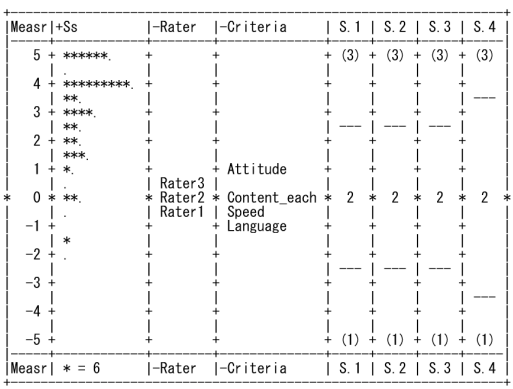
図1.

変数マップ

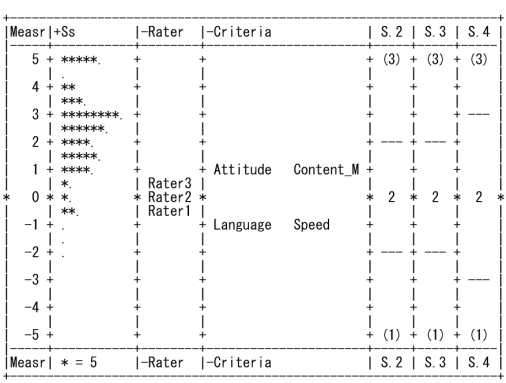
ディスカッション



ディベート(内容個人点)



ディベート(内容グループ点)



注. Ss = 受験者。Rater = 採点者。Criteria = 観点。S.1〜S.4 = ルーブリックの1〜4個目の観点(表4参照)。Speed = 音量等。ディベート(内容グループ点)のS.1がないのは、グループの平均値を使ったことでレベル1の採点がなくなり、2段階になったため(付表G)。

表9.
採点者、受験者、ルーブリック観点の統計値

	測定値の平均値	測定値のSD	標準誤差の平均値	適合度	層(Separationと Strata)		信頼性
ディスカッション							
受験者	2.48	2.64	1.16	範囲外あり	1.91	2.88	.79
採点者	0.00	0.14	0.13	範囲内	0.42	0.90	.15
観点	0.00	0.73	0.12	範囲内	5.87	8.16	.97
ディベート(内容個人点)							
受験者	3.02	1.87	0.97	範囲外あり	1.50	2.33	.69
採点者	0.00	0.29	0.10	範囲内	2.66	3.88	.88
観点	0.00	0.77	0.12	範囲内	6.49	8.99	.98
ディベート(内容グループ点)							
受験者	2.43	1.58	0.95	範囲外あり	1.20	1.94	.59
採点者	0.00	0.30	0.10	範囲内	2.85	4.14	.89
観点	0.00	0.94	0.12	範囲内	8.02	11.03	.98

注. 表の見方は、表1参照。信頼性は分離信頼性。受験者と観点では、高い値がそれぞれを詳細に分けて測れることを意味し、高い値が望ましい。一方採点者では、高い値は採点者の厳しさが大きく異なることを意味するため、低い値が望ましい。

受験者と観点

受験者の適合度に関して、ディスカッションとディベートともに範囲外の適合しない者がおり、過剰適合と不適合が両方あった。例えば、Infit MS 0.49以下の者が10.13～11.45%、1.51以上の者が10.13～12.33%、2.01以上の者が2.64～7.05%だった（表9参照）。受験者のスコア（回答）がラッシュ分析による予想パターンに似すぎていたり（過剰適合）、予想から大きく異なっていたり（不適合）した受験者がある程度いたということである。測定の質を下げる可能性がある不適合の詳細を、予想外の回答（Unexpected responses）のパターンで調べたところ、採点者間で異なるパターンはほとんどなく、ルーブリックのある観点のスコアが他観点のスコアから予想されるスコアよりも高いか低く、そのパターンが予想と異なるものだったことに起因していた。例えばディスカッションでは、技術点のスコアが予想より高い場合（40.91%）と内容のスコアが予想より低い場合（36.36%）が多かった。ディベート（内容個人点）では、態度または言語のスコアが予想より低い場合が多く（30.00～40.00%）、ディベート（内容グループ点）では、内容と言語のスコアが予想より低い場合が多かった（30.00～45.00%）。ディベート（内容グループ点）の内容で予想外と判定されたスコアは、個人の力だけでなく、グループの他のメンバーの力で変わるため、個人の力の反映である他の観点とのずれが見られたためだろう。全体的には、受験者の不適合は、スピーキング力の構成要素の相対的な高低によって起きる場合が多いと推測できる。これは、ある観点が極端に苦手や得意などの、もともと持つ力の高低差とともに、グループでの役割や構成、テストを受ける順序など、テストの環境や受験者の情意面の影響も考えられる。Bonk and Ockey（2003）は、受験者の不適合はSTではあまり大きな問題ではないと述べ、その理由を、(a) STでは当て推量で答える、真面目に取り組まない、寝ているような可能性は低く、受験者の採点が採点者2名から出ているために結果が安定しないことから起こることが多いため、(b) ルーブリック観点の難易度は全受験者から計算しているが、それとは異なる、観点ごとの得意不得意が個々の受験者に存在することは十分あるため、と述べている。本研究では、全員が真剣にテストに取り組んだことは採点者が確認しており、(b) の事象が多いことは上述の予想外の回答の分析で述べた。そのため、受験者の不適合の多さはあまり問題にならないと考えた。MFRMの受験者適合度基準は厳しすぎる場合があり、Mokken尺度分析（Walker & Wind, 2020）や、タスク数やサンプルサイズを考慮したBootstapping法（Seol, 2016）が適切な場合も報告されており、今後は複数の基準で確認する方向もあるだろう。

受験者がSTで弁別できるか（分離）については、Strataを見ると、ディスカッションで2.88で、能力層が異なる2～3群に分けられていた。ディベートでは2.33, 1.94であり、2群に分けられる程度だった。受験者分離信頼性は、一般的な分析でのテスト信頼性と概念的に同じで、ディスカッションとディベート（内容個人点）では .79と .69とまずまずの高さだったが、ディベート（内容グループ点）では .59であり、テスト信頼性は低いという結果だった。これは、ディベート（内容グループ点）の4観点中の3観点において、採点でレベル1がほぼ使われず、受験者の能力を3段階で弁別することができなかったことが大きな理由と思われる（付表G参照）。

観点については、適合度はすべて範囲内におさまっていた。分離については、ディスカッションのStrataで8.16あり、意図した9段階（3観点の3段階ずつのルーブリック）まではいかないものの、それに近いものが得られており、ディスカッションでも全体的には同様だった。観点の分離信頼性は、.97～.98と高かった。

ルーブリックの適切さについては、Bond et al. (2021)のほとんどの基準を満たしていた(詳細は付表G)。満たしていなかった基準は、(x) 各レベル使用頻度が10以上という基準で、レベル1の頻度が極端に少なかった点と、(y) 隣接する敷居値の距離が1.40〜5.00という基準において、5.01以上が見られ、レベル2の距離が長く(レベル2をとった者が多く)、レベル1とレベル3の難易度の差が大きすぎた点だった。この点を今後修正すべきかを検討する際には、レベル1と2、またレベル2と3の受験者のスコアや発話を比較して、本来レベル1(レベル2)になる受験者がレベル2(レベル3)に入っていないか(またその逆もないか)、それがある場合に、レベル1から2になるのをより難しくするか、レベル2から3になるのをより易しくするか、レベル2を分割して、3段階を4段階にするか、問題があった箇所の記述やサンプル例を明確化したり修正したりするか、採点者トレーニングでの説明を変えるかなどを1つずつ検討していくことになる。もちろん授業目標や授業内容、授業での反応とルーブリック結果を比較し、意図通りであれば修正しない方向もある。例えば (x) の場合、レベル2は授業目標をおおむね満たした場合であり、指導の結果、それを全生徒が到達した場合には、レベル1の使用数は0回になるが、それは問題がないと考えられる。また (y) の場合に意図的にレベル2の幅を広くしたときには問題ないだろう。また観点の段階数を増やすことで採点の負担も増すため、それも考慮すべきである。

採点者

採点者信頼性に関して、採点者の厳しさ値の違いは小さかった(表10参照)。例えばディスカッションでは、厳しさ値の差の最大値は0.34(-0.18〜0.16)であり、fair scoreの差の最大値は3点満点中0.07のみだった。表9で採点者信頼性は、.15と低かった。この値は、採点者がどの程度異なる採点をしたかを示す分離信頼性であり、低い値は採点者が似た厳しさで採点していたことを示し、今回はそうだったと言える。採点者間の一致率は高く、また予測一致率よりも高かった(72.9% > 71.6%)。他のテストでも同様の結果だった。

表10.
採点者統計値

	厳しさ値の差の最大値	fair scoreの差の最大値	Infit MS	Outfit MS	一致率	予測一致率
ディスカッション	0.34	0.07	0.86〜1.35	0.80〜1.41	72.9%	71.6%
ディベート(内容個人点)	0.69	0.16	0.93〜1.09	0.93〜1.11	82.9%	67.6%
ディベート(内容グループ点)	0.72	0.14	0.95〜1.02	0.97〜1.05	81.6%	66.7%

採点者適合度のInfit MSとOutfit MSの点では、全員の採点者が0.50〜1.50内に入り、適合度を満たしていた。例えばディスカッションでは、0.86〜1.35と0.80〜1.41だった。この結果から、本研究のテスト結果では、採点者内一貫性が高く、採点者が最初から最後まで、また観点ごとに似た厳しさで採点していたことが分かる。

一般化可能性理論（G theory）

G theoryのG研究の結果に基づく各変動要因の分散要因の割合を表11に示す。例えばディスカッションの内容の観点（C1）では受験者が62.00%、採点者が0.00%、残差が38.00%だった。残差は、採点者が受験者によって採点方法を変えたことから起きる変動と、それでは説明できない変動を含む値である。ここから、受験者の分散が占める割合が大きく、採点者の影響はなかったことが分かる。表11全体で採点者の影響は最大で2.14%で、小さかったことが示された。

次に、採点者数を変えた時にテストの信頼性がどのように変わるかをD研究で調べた。表12によると、例えばディスカッションの内容（C1）では、採点者1名だと .62で .70未満のため十分な信頼性がなく、2名いると .77となり、必要な信頼性を満たしていた。表現（C2）、技術点（C3）でも同様で、全体（計）では採点者1名で十分だった。ディベートでもほぼ同じ結果だった。

表11.
G研究における推定された分散成分とその割合

		ディスカッション			ディベート（内容個人点）				ディベート（内容グループ点）			
		C1	C2	C3	C1	C2	C3	C4	C1	C2	C3	C4
受験者 (p)	VC	0.23	0.23	0.23	0.22	0.23	0.15	0.15	0.10	0.23	0.15	0.15
	%	62.00	66.94	59.85	74.38	67.87	65.90	56.04	49.92	67.87	65.90	100.00
採点者 (r)	VC	0.00[a]	0.00[b]	0.00[b]	0.00[b]	0.01	0.00[b]	0.00[a]	0.00[b]	0.01	0.00[b]	0.00[a]
	%	0.00	0.09	0.08	0.75	1.89	1.42	0.00	2.14	1.89	1.42	0.00
残差 (pr, e)	VC	0.14	0.11	0.15	0.07	0.10	0.07	0.11	0.10	0.10	0.07	0.00
	%	38.00	32.98	40.07	24.87	30.25	32.68	43.96	47.94	30.25	32.68	0.00

Note. C1～C4 = 第1～4の観点。VC = 分散要因（variance component）
[a] 負の分散を0に固定。[b] VCの小数点第3位を四捨五入して0.00になった。下の段の%はVCの四捨五入前の値を用いて%を算出した。

表12.
D研究における信頼性の変化

	ディスカッション				ディベート（内容個人点）					ディベート（内容グループ点）				
R	C1	C2	C3	計	C1	C2	C3	C4	計	C1	C2	C3	C4	計
1	.62	.67	.59	.81	.74	.68	.66	.56	.79	.50	.68	.66	.56	.71
2	.77	.80	.74	.89	.85	.81	.79	.72	.88	.67	.81	.79	.72	.83
3	.83	.86	.81	.93	.90	.86	.85	.79	.92	.75	.86	.85	.79	.88

注. R = 採点者。計 = 3～4観点を合わせた結果。下線はΦ = .70 以上の場合。

テスト間の相関

STの実施や採点には手間がかかるため、学期で複数回行うべきかが議論になることがある。その点への示唆を得るために、MFRMで出した受験者能力値の相関を調べたところ、ディスカッションとディベート間では中程度の相関があった（r = .41, .46）。またディベートの内容個人点と内容グループ点の相関は非常に高く（r = .94）、どちらを用いても、受験者のスコアは似た結果となった。

ディスカッションとディベートの間では中程度しか相関がなく、共通して測れる部分は2乗した16.81～21.16%のみであった。9月と11月で2カ月の間なので、能力の違いよりは、テスト形式の違いによるスピーキング力の見え方の違いが大きいと解釈でき、1形式だけでテストを行うと測れない力も多いと考えられる。

考察

研究課題1の採点者間一致度・一貫性の点での採点者信頼性については、全体的に保たれていた。MFRMでの採点者の厳しさ値の差は小さく、一致度も十分あった（表10）。G theoryのG研究でも、採点者がテストのスコア全体に影響する割合は小さかった（表11）。しかし採点者の組み合わせごとにみると、kappa係数や相関係数に低いものがあり（表6）、その点を満たすためには採点者トレーニング等が必要だと思われる。

研究課題2の採点者内一貫性の点については、MFRMの採点者適合度は範囲内で、保たれていた（表10）。

研究課題3の必要な信頼性を満たすための採点者数（表12）については、G theoryのD研究結果から、3～4観点の合計点では採点者1名で十分と示された。1観点のスコアで何らかの判断をする場合には、必要な採点者数は1～3名と分かった。

本結果とK&Wのグループ型ディスカッションの結果を比較すると、採点者信頼性の点で共通点は多かった。例えばMFRMの採点者の厳しさ値にはあまり違いがなく、一致度が高かったが、採点者ごとの組み合わせでの一致度と一貫性を見ると低いものも見られた点が同じだった。一方大きく異なる点として、グループ型タスクを使った合計点の結果で、必要な採点者数が、K&Wでは3または4名だったが本研究では1名だった。2つの研究は、シンプルなルーブリックを使って、詳細な採点者トレーニングはない状態で採点という点が共通だが、相違点も多く、以下に7点挙げる。

第1に、生徒の学力層が異なっていた。本研究は高校3年生対象で、その英語力は県の上位層に属していたが、K&Wの高校1年生の英語力は県で平均的だった。

第2に、ST結果を成績点に含めるかが異なり、本研究では評定の3割を占めていたが、K&Wでは形成的評価のみで成績には含めなかった。

第3に、受験者の能力と比較したテストの難易度は、図1で受験者の多くが上の方に位置づけられたことから分かるように本研究では易しめだった。K&Wでは2回行ったグループ型ディスカッションのうち2回目は本研究と同様だったが、1回目は適した難易度の範囲だった。

第4に、ルーブリックの観点の違いがあった。本研究では、内容と言語的要素、コミュニケーションに対する姿勢を採点対象としたが、K&Wでは、内容とやり取りの適切

さ、コミュニケーションへの意欲を対象とし、言語面でなく「やり取りの適切さ」を採点し、やり取りの適切さの判断の難しさが採点の不安定さにつながった可能性がある。

第5に、生徒1人あたりの発言時間が異なっていた。本研究のディスカッションでは、1グループ約5分、1人あたり75秒以上あったが、K&Wでは1グループ3〜4分、1人あたり45秒から60秒で、グループ全員が十分話す時間がなく、採点者がスコアの結論を出す前に採点の時間が終わってしまうこともあった。Van Moere（2006）では1グループ4名で5〜10分採点にかかったとあり、自由に話す形で採点者信頼性を保つためには1グループ採点により長い時間を確保する必要があると思われる。

第6に、採点者の認知的負荷の度合いが違ったと思われる。本研究では、司会者と参加者を決め、参加者が発言する順番も決まっていたため、会話の流れは予測できた。K&Wでは全員が自由に話してもいい形だったため、誰がいつ話すか予想ができず、採点者は生徒Aが話し始めるとその発話を採点し、スコアの結論が出ない状態でも、次の生徒Bが話し始めるとBの採点に移り、生徒Aが再び話し始めると前回のところから再開してスコアを考えた。1グループの4人分を同時に採点する形で、採点がより難しくなっていたと考えられる。

第7に、採点者である教員の特徴が異なっていた。本研究では、採点者全員が同じ学年に対して同じ教材を用いて指導や評価を行い、指導理念を共有していた。そのため、直接教えていない生徒の発話の採点でも、自分が教える生徒の様子からスピーキング力がどのくらいかを判断しやすかった可能性がある。さらに本研究のSTの前の6月に2種類のSTを実施し、また5年前から1科目ごとに年4回STを行っていたため、ST実施・採点の経験もあった。一方K&Wでは、授業担当者以外の教員は、STを行っておらず、授業中のスピーキング活動内容もあまり共有されていない中で、他教員が採点の援助をする形で、また授業担当者も前年にはSTを実施していなかった。そのため、どの程度の発話を求めるか等の共通認識を持っておらず、判断がずれた可能性がある。

本研究とK&Wの相違点の中で、どの点の影響が大きかったかについては要因を絞って今後実証研究を行う必要があるが、第4〜7の相違点での説明は理解しやすいものである。シンプルなルーブリックで採点者トレーニングなしのグループ型タスクでも、やり取りの適切さよりは発話の言語面（文法や語法）に注目して採点する、1グループあたりの時間を長めにとる、生徒の発言する順番を決めるなど、会話の流れがある程度予想できるタスク構造にする、指導や評価について共通認識を持つ教員が採点するようにすることで信頼性が保てる可能性がある。

ただし、これらの方法は信頼性を高める可能性がある反面、妥当性や実行可能性を下げることも認識しておきたい。例えば、会話での役割や発言の順序を決めることは、グループ型ディスカッションで測りやすいとされる、やり取り力を測る度合いが下がる。役割や発言順番が決まっていないことで、会話を開始したり、会話を修復したりという発話が生徒から自然な形で自発的に表れにくくなるためである。さらにルーブリックの観点にやり取りの適切さを含めないことで、やり取り力の重要な要素を測りにくくなる。また1グループあたりの時間を長く確保することで、テストに必要な授業時間が増え、その分指導時間は減るため、他教員の同意が取りにくくなり、共通認識のある教員に採点を依頼することも難しくなりやすい。このような点も含めて、どのようなタスク形式や実施方法にするかを決める必要がある。

本研究とK&Wの結果により、授業内STでは、シンプルなルーブリックを使えば、長時間の採点者トレーニングは必要なく、採点者1名でも安定して採点できる場合もあることが示された。しかし、表7にあるように、全体として信頼性が確保できていても、個々に見ると一致度や一貫性が低い場合もあり、3段階のレベル1と2や、レベル1と3と意見が割れた場合も見られた。そのような場合、素点を用いて採点者1名のときには、採点の偏りがそのまま成績に直接反映されることになる。そのため、一般にSTの採点で言われているように、事前の採点者トレーニングを短時間でも行い、特にずれそうな点だけでも行うことが望ましい。またK&Wで行ったように、テストの最初の数件を採点した後に話し合いの機会が持てるとよいだろう。実際のSTでの生徒の発話を観察すると、ルーブリック等のあいまいな点等に気づき、基準を調整した上で採点を行うことができる。

次に、本研究でのグループ型のディスカッションとディベートを比較すると、採点者信頼性の点でも、他の点でもテストの性質に関わる相違点は少なかった。また2つのタスク形式から推測できる受験者の能力値の相関は中程度であり、形式は似ていても測っている力は同じではないことも示された（***テスト間の相関参照***）。K&Wを再分析すると、1〜3カ月の間をおいた時期が隣り合わせのテスト同士では、弱から中程度の相関（r = .31〜 .40）があり、同様の傾向があった。もし2回STを行ううちのどちらかだけでSTに関する評定を決定すると、実施したテスト形式を得意とする生徒が有利になる可能性がある。使用タスクによって見える力が異なることは先行研究でも述べられており（In'nami & Koizumi, 2016; Ockey et al., 2015）、定期的に様々なタスクを用いてテストを実施する方が望ましいだろう。

ディベートで内容個人点と内容グループ点を使った場合を比較すると、内容個人点の方が生徒の個人の力がより反映されるため、より受験者信頼性が高く（表9）、受験者の不適合は少なかった（例［表9には不掲載］：Infit MS 1.51以上が、内容個人点の場合に12.33%、内容グループ点で14.10%）。内容グループ点の不適合では、内容のスコアが予想より低いというケースが多く（45.00%）、内容にグループ全体の力が反映されているとはいえ、STの信頼性の点では内容個人点を使用した方がより適切と言えよう。しかし、グループで協力して相手グループを論破するというディベートの目的と教育的効果を考えると、ルーブリックでもそれが反映されていた方がよく、その点で内容をグループ点にする方法は適切だと考えられる。内容グループ点を使った結果は、STの信頼性は若干下がる傾向になることを理解しつつ使うのがよいだろう。

結論

本研究は、3〜4観点で3段階のシンプルなルーブリックを用いることで、詳細な採点者トレーニングがない場合にも信頼性が十分保てるかを、グループ型のディスカッションとディベートの場合で検証した。その結果、採点者間一致度・一貫性と採点者内一貫性の観点で十分な信頼性が満たされていることが示された。K&Wとの比較で、信頼性を高める方法として、グループ型タスクの場合に、やり取りの適切さよりは言語面（文法・語法）を観点に入れる、生徒が話す時間を長めに確保する、生徒の役割や発言する順番を決める、共通認識がある教員で採点を行う等の方法が示唆された。

本研究で得られた示唆は4点ある。第1に、採点者トレーニングを詳細に行えない状況でも採点者信頼性を保つ方法として、タスク構造や手順の工夫などが示唆さ

れ、教員が授業内STを作成・実施・採点する際の指針として使えると思われる。特にK&Wに加え、新たなタスク形式（ディベート）や生徒の実態が異なる別の学校での結果が報告されたことで、より適用できる範囲が増える可能性がある。第2に、ディベートの内容の観点について、個人点とグループ点を用いることの利点と弱点が示され、グループで協力して行う度合いが大きいタスクにおいてどちらのスコアを用いるかを考える際に役立つだろう。第3に、近い時期に行ったSTスコアの関係は中程度のみと示されたことで、異なるテスト形式を用いることで多様な生徒の力を的確に測ることの意義が示された。第4に方法論的な示唆として、採点者信頼性を報告する際に、一致度・一貫性・測定アプローチの1つだけを用いることが多いが、それぞれの指標で結果が若干異なる事例を本研究は報告した。目的によるが、可能であれば3つのアプローチすべての情報を提供することで、バランスよく採点者信頼性を検討できるだろう。またその分析によって、採点者間や採点者内でスコアの大きなずれが見られた場合を特定することができ、録音・録画を使ってなぜずれたのかを確認し、採点でずれやすい観点や特徴、状況を精査し、それ以降の採点者トレーニングで共有するという流れを作ることもできる。

本結果は、今回の使用したタスク形式やルーブリック、指導目標や指導タスク、生徒の英語の熟達度、生徒と教員の情意的や認知的特徴などに限定される可能性がある。そのため、別な状況での研究をさらに行い、知見を深めることが必要である。要因を厳密に統制した実験的な研究は授業内STの研究は行いにくいが、より多くの研究を行うことで、日本の状況に適したSTの実施や採点の方法を、発展させていくことにつながるだろう。

最後に、採点者信頼性が満たされたとしても、測りたい力が適切に測れ、テストの目的に沿った形で適切に使えているか、という妥当性の中の一部が満たされただけである。例えば論証に基づく妥当性検証を、領域定義、得点化、一般化、説明、外挿、利用、波及効果という7段階の推論に基づくものと捉えるとき、採点者信頼性は、得点化、一般化に大きく関わる（Chapelle & Voss, 2021; 小泉, 2018b）。教室内STのスコアに基づく解釈と使用の妥当性を示すためには、採点者信頼性（得点化、一般化）以外の観点を検証する必要があり、様々な文脈での教室内STにおいて、その検討を教員と研究者が協力しながら行うことが重要である（Gu, 2020; Koizumi, 2022）。

注

1. 念のため信頼性以外の値の検討のために重みづけを行ったデータを用いて、ディベートのデータのMFRMを行ったところ、本研究で報告した結果と全体的には変わらなかった（観点に重みをつけて分析する方法は、Linacre, 2021, pp. 375–377参照）。相違点は、重みづけがあることで、受験者・採点者・観点すべてにおいて層（SeparationとStrata）が増大し、信頼性が高くなっていたことだった。例えば、ディベート（内容個人点）のStrataでは、受験者が6.79、採点者が11.89、観点が22.73であり、表9の2.33, 3.88, 8.99よりも大きくなっていた。また表8で報告した全体的モデル適合度は、重みづけがあることでやや適合しない結果になった（例：ディベート［内容個人点）の重みづけありで5.54%, 1.66%）。この結果から、受験者の力を詳細に弁別したいのであれば観点を重みづけるのは1つの方法ではあるが、採点者の弁別も高まり（採点の厳しさの違いも大きくなり）、モデルに適

合しにくくなる場合もあるため、重みづけを行う場合にはそれも考慮に入れるべきであろう。

謝辞

本研究は、科研費基盤研究（C）20K00894と、宮城県教育委員会主催令和元年度「発信型英語教育拠点校事業」（文部科学省事業「英語教育改善プラン推進事業生徒の発信力強化のための英語指導力向上事業）の助成を受けたものである。ご協力くださった渡邉聡代先生、先生方、生徒の皆さん、重要なご指摘をくださった2名の査読者に感謝申し上げたい。

著者略歴

小泉利恵は清泉女子大学教授である。主な研究対象は言語テスティングのスピーキング評価で、学習のための評価の実現に関わる諸問題を扱っている。

初澤晋・磯部礼奈は宮城県石巻高校の、**松岡京**一は宮城県築館高校教諭である。宮城県内の公立高校で指導し、英語の指導と評価の実践・研究を重ねてきている。

引用文献

小泉利恵（2018a）「一般化可能性理論」平井明代編『教育・心理・言語系研究のためのデータ分析―研究の幅を広げる統計手法』（pp. 65–93）東京図書.

小泉利恵（2018b）『英語4技能テストの選び方と使い方―妥当性の観点から―』アルク.

小泉利恵（2018c）「ダイアローグ型タスクによるスピーキング能力評価研究の動向―学習者対話型テストを中心に―」 In S. Ishikawa (Ed.), *Learner Corpus Studies in Asia and the World: Vol. 3 Papers from LCSAW2017* (pp. 27–42). Kobe University. http://www.lib.kobe-u.ac.jp/kernel/seika/ISSN=21876746.html

嶋田和成（2017）「相関分析―変数間の関係を分析する」 平井明代 (編著)『教育・心理系研究のためのデータ分析入門―理論と実践から学ぶSPSS活用法』（第2版, pp. 145–164）東京図書.

髙島健治（2019）『思考力を働かせる学習サイクルの構築―定型表現で終わらない［やり取り］の力を育成するために―』平成30年度千葉県長期研修研究報告書.

投野由紀夫・根岸雅史編（2020）『教材・テスト作成のためのCEFR-Jリソースブック』大修館書店.

平井明代・横内裕一郎・加藤剛史（2018）「項目応答理論：標本依存と項目依存を克服した測定を実現する」平井明代編『教育・心理・言語系研究のためのデータ分析―研究の幅を広げる統計手法』（pp. 94–137）東京図書.

福田純也・田村祐・栗田朱莉（2017）「中学校教科書における口頭コミュニケーションを志向した活動の分析―第二言語習得研究におけるタスク基準からの逸脱に焦点をあてて―」 *JALT Journal*, *39*(2), 165–182. https://doi.org/10.37546/JALTJJ39.2-4

Benesse Corporation. (2019).「GTEC英語教育に関する調査結果」https://www.benesse.co.jp/gtec/schoolofficials/research/

松尾美幸(2019)「事例報告　テストが到達目標と指導に与える影響」https://www.britishcouncil.jp/programmes/english-education/japan/report/assessment2018-seminar/case1

松村香奈・守屋亮(2019)「教室におけるPaired oral testの診断的評価および学習者の受容に関する調査―混合研究法を用いて―」*EIKEN BULLETIN*, *31*, 212–234. https://www.eiken.or.jp/center_for_research/list_1X/

文部科学省(2019)「平成29・30・31年改訂学習指導要領(本文、解説)」https://www.mext.go.jp/a_menu/shotou/new-cs/1384661.htm

文部科学省(2020)「令和元年度『英語教育実施状況調査』の結果について」https://www.mext.go.jp/a_menu/kokusai/gaikokugo/1415042.htm

Akiyama, T. (2001). The application of G-theory and IRT in the analysis of data from speaking tests administered in a classroom context. *Melbourne Papers in Language Testing*, *10*, 1–21. http://wayback.archive-it.org/1148/20130404065508/http://ltrc.unimelb.edu.au/mplt

Akiyama, T. (2004). *Introducing EFL speaking tests into a Japanese senior high school entrance examination* [Unpublished doctoral dissertation]. University of Melbourne.

Aso, Y. (2000). A comparison of holistic and analytic scorings for oral interview tests. *ARELE*, *11*, 131–139. https://doi.org/10.20581 /arele.11.0_131

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge. https://doi.org/10.4324/9781315814698

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89–110. https://doi.org/10.1191/0265532203lt245oa

Celce-Murcia, M. (2007). Rethinking the role of communicative competence in language teaching. In E. Alcón Soler & M. P. Safont Jordà (Eds.), *Intercultural language use and language learning* (pp. 41–57). Springer Netherlands. https://canvas.harvard.edu/files/926812/download?download_frd=1&verifier=HL5njGKyAX7HvsYMG1lDrU3H57BhuU4dLI4qrELT

Chapelle, C. A., & Voss, E. (Eds.). (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge University Press.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, *26*(3), 367–396. https://doi.org/10.1177/0265532209104667

Ellis, R., & Shintani, N. (2014). *Exploring language pedagogy through second language acquisition research*. Routledge. https://doi.org/10.4324/9780203796580

Engelhard, Jr., G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge. https://doi.org/10.4324/9780203073636

Galaczi, E. D., & Taylor, L. B. (2021). Measuring interactional competence. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 338–358). Routledge. https://doi.org/10.4324/9781351034784

Grabowski, K. C., & Lin, R. (2019). Multivariate generalizability theory in language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment* (Vol. I, pp. 54–80). Routledge. https://doi.org/10.4324/9781315187815

Gu, P. Y. (2020, October 30–31). *Validity in classroom-based formative assessment* [Keynote speech]. New Directions 2020 (online). https://www.youtube.com/watch?v=EHxULkiVC9I&t=14s

Hirai, A., & Koizumi, R. (2013). Validation of empirically derived rating scales for a Story Retelling Speaking Test. *Language Assessment Quarterly*, *10*(4), 398–422. https://doi.org/10.1080/15434303.2013.824973

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing*, *33*(3), 341–366. https://doi.org/10.1177/0265532215587390

Inoue, C. (2013). *Task equivalence in speaking tests.* Peter Lang.

Iwamoto, N. (2018). An investigation of Japanese university students' English speaking skills.『日本英語英文学会 英語学論説資料集』, *50*(6分冊増刊), 346–353. http://www.jaell.org/gakkaishi26th/Iwamoto.pdf

Iwashita, N., May, L., & Moore, P. (2021). Operationalising interactional competence in computer-mediated speaking tests. In R. Salaberry & A. R. Burch (Eds.), *Assessing speaking in context: Expanding the construct and its applications* (pp. 283–302). Multilingual Matters.

Jönsson, A., Balan, A., & Hartell, E. (2021). Analytic or holistic? A study about how to increase the agreement in teachers' grading. *Assessment in Education: Principles, Policy & Practice*, *28*(3), 212–227. https://doi.org/10.1080/0969594X.2021.1884041

Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: Issues, options and directions*. Equinox.

Koizumi, R. (2022). L2 speaking assessment in secondary school classrooms in Japan. *Language Assessment Quarterly*, *19*(2), 142–161. https://doi.org/10.1080/15434303.2021.2023542

Koizumi, R., In'nami, Y., & Fukazawa, M. (2020). Comparison between holistic and analytic rubrics of a paired oral test. *JLTA Journal, 23*, 57–77. https://doi.org/10.20622/jltajournal.23.0_57

Koizumi, R., & Watanabe, A. (2021). Rater reliability in classroom speaking assessment in a Japanese senior high school. *ARELE*, *31*, 129–144. https://doi.org/10.20581/arele.32.0_129

Lin, C.-K. (2017). Working with sparse data in rated language tests: Generalizability theory applications. *Language Testing*, *34*(2), 271–289. https://doi.org/10.1177/0265532216638890

Linacre, J. M. (2021). Facets: Many-facet Rasch measurement (Version 3.83.6) [Computer software]. MESA Press. https://www.winsteps.com/facets.htm

Linacre, J. M. (2021). *A user's guide to FACETS Rasch-model computer programs: Program manual 3.83.5*. http://www.winsteps.com/manuals.htm

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

McDonald, K. (2018). Post hoc evaluation of analytic rating scales for improved functioning in the assessment of interactive L2 speaking ability. *Language Testing in Asia*, *8*(19), 1–23. https://doi.org/10.1186/s40468-018-0074-3

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

McKay, T. H., & Plonsky, L. (2021). Reliability analyses: Estimating error. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 467–482). Routledge. https://doi.org/10.4324/9781351034784

McNamara, T. F. (1997). "Interaction" in second language performance assessment: Whose performance? *Applied linguistics, 18*(4), 446–466. https://doi.org/10.1093/applin/18.4.446

McNamara, T., Knoch, T., & Fan, J. (2019). *Fairness, justice, and language assessment*. Oxford University Press.

Mizumoto, A. (2021). *Cohen's kappa and other interrater agreement measures*. http://langtest.jp/shiny/kappa/

Nakatsuhara, F. (2007). Developing a rating scale to assess English speaking skills of Japanese upper-secondary students. *Essex Graduate Student Papers in Language & Linguistics*, *9*, 83–103. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.144.8523&rep=rep1&type=pdf

Negishi, J. (2011). *Characteristics of group oral interactions performed by Japanese learners of English*. (Publication No. 5722) [Doctoral dissertation, Waseda University]. Waseda University Repository. http://hdl.handle.net/2065/37662

Negishi, J. (2015). Effects of test types and interlocutors' proficiency on oral performance assessment. *ARELE*, *26*, 333–348. https://doi.org/10.20581/arele.26.0_333

Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing*, *31*(2), 147–175. https://doi.org/10.1177/0265532213514401

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing*, *32*(1), 39–62. https://doi.org/10.1177/0265532214538014

Plonsky, L., & Mizumoto, A. (2021). *Alternative reliability indexes*. https://lukeplonsky.shinyapps.io/omega/

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, *29*(2), 223–241. https://doi.org/10.1177/0265532211421162

Seol, H. (2016). Using the bootstrap method to evaluate the critical range of misfit for polytomous Rasch fit statistics. *Psychological Reports*, *118*(3), 937–956. https://doi.org/10.1177/0033294116649434

Stapleton, P., & Collett, P. (2010). *JALT Journal* turns 30: A retrospective look at the first three decades. *JALT Journal*, *32*(1), 75–90. https://jalt-publications.org/jj/issues/2010-05_32.1

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation*, *9*(4), 1–11. https://doi.org/10.7275/96jp-xz07

University of Iowa, College of Education. (出版年不明). Computer programs. https://education.uiowa. edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing*, *23*(4), 411–440. https://doi.org/10.1191/0265532206lt336oa

Walker, A. A., & Wind, S. A. (2020). Identifying misfitting achievement estimates in performance assessments: An illustration using Rasch and Mokken scale analyses. *International Journal of Testing*, *20*(3), 231–251. https://doi.org/10.1080/15305058.2019.1673758

Wells, C. S., & Wollack, J. A. (2003). *An instructor's guide to understanding test reliability*. University of Wisconsin. https://testing.wisc.edu/instructionalsupport.html

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370. https://www.rasch.org/rmt/rmt83b.htm

Yokouchi, Y. (2018). *Effects of task conditions on spoken performance in retelling*. (Publication No. DA08786) [Doctoral dissertation, University of Tsukuba]. University of Tsukuba Repository. https://doi.org/10.15068/00153779

付表A

ディスカッションのルーブリック

Performance Test Evaluation Sheet: English Expression II
[司会者:Moderator]
Class: No. Name

内容	まとめ方	参加者の発言を別の表現で言い換え、要約できている。	10
		参加者の発言を繰り返すなどして、適宜フィードバックしている。	7
		参加者の発言に言及しようとしているが、不十分である。	4
表現	発声・音声・ノンバーバルコミュニケーション	適切にアイコンタクトやジェスチャーなどを用いており、自然に議論を進行することができている。	10
		時折アイコンタクトやジェスチャーを用いており、自然な議論の進行に努めている。	7
		アイコンタクトやジェスチャーが不十分であり、議論の進行に滞りがある。	4
技術点	文法・語法	文法・語法のエラーが少なく、既習の表現を活用しながら自分の考えを不足なく伝えられている。	10
		文法・語法のエラーについて多少のエラーや単純な表現は多いが、自分の考えが伝えられている。	7
		文法・語法のエラーが目立ち、自分の考えが十分に伝えられていない。	4

[参加者:Participant]

Class:	No.	Name	
内容	考え方	相手の考えと自分の考えを比較しながら、根拠に基づいた自分の考えを述べている。	10
		根拠に基づいた自分の考えを分かりやすく述べている。	7
		自分の考えを述べているが、内容がやや不足である。	4
表現	発声・音声・ノンバーバルコミュニケーション	適切にアイコンタクトやジェスチャーなどを用いており、自然な(理解されやすい)表現をしている。	10
		時折アイコンタクトやジェスチャーを用いており、自然な(理解されやすい)表現に努めている。	7
		アイコンタクトやジェスチャーが不十分であり、相手に伝わりにくい表現である。	4
技術点	文法・語法	文法・語法のエラーが少なく、既習の表現を活用しながら自分の考えを不足なく伝えられている。	10
		文法・語法のエラーについて多少のエラーや単純な表現は多いが、自分の考えが伝えられている。	7
		文法・語法のエラーが目立ち、自分の考えが十分に伝えられていない。	4

付表B

ディベートのルーブリック

Performance Test Evaluation Sheet: Communication English III

(内容は各立場で異なり、他の観点は共通)

【肯定側立論／否定側立論】(Affirmative Constructive Speech / Negative Constructive Speech)

Class:	No.	Name
内容	12	トピックについて、自分の立場に即して考えや意見を理由や具体例を添えて分かりやすく説明している。
	8	トピックについて、自分の立場に即して考えや意見を説明している。
	4	トピックについて、自分の立場に即して考えや意見を説明しようとしているが、内容が不明瞭である。

コミュ姿勢	8	自然なアイコンタクトやジェスチャーを用いている。積極的に発言して議論の発展に努めている。
	5	アイコンタクトやジェスチャーを用いる努力がある。時々言いよどみながらも議論の発展に努めている。
	3	アイコンタクトやジェスチャーが不十分である。積極性に欠け、議論への参加に意欲が見えない。
文法語法	5	理解を妨げるような文法・語法のエラーがほぼなく、自分の考えを不足なく伝えられている。
	3	理解を妨げるような文法・語法のエラーが少しあり、自分の考えが十分に伝わらないところがある。
	1	理解を妨げるような文法・語法のエラーが多く、自分の考えが十分に伝えられていないところが多い。
音量速度発音	5	話す音量や速さを調整し、また発音にも留意しながら、聞き手が理解しやすいように話している。
	3	話す音量や速さを調整し、聞き手が理解しやすいように努めている。
	1	話す音量や速さが、聞き手にとって理解しづらいものである。
12点	We think that Japanese high schools should allow their students to work part time. Some need to earn money to pay school fees, some to buy essential goods for their school life, and others to support their families financially.	
8点	We think that Japanese high schools should allow their students to work part time. Students can learn many things through working part time. The experience of part-time job will be helpful for their future.（具体性に欠ける）	

【質疑】(Question)

内容	12	相手側の立論で提起した内容について質問している。
	8	相手側の立論の内容について質問しようとしているが、内容が不明瞭である。
	4	自然なアイコンタクトやジェスチャーを用いている。積極的に発言して議論の発展に努めている。
12点	You said that some students need to support their family financially so they should be allowed to work part time. I know what you mean, but don't you think it will have a negative effect on their study? They will definitely have less time to study.	
8点	I see your point. But we're afraid that they will study less due to their part-time job. What do you think?（要約なし）	

【反論】(Rebuttal Speech)

内容	12	立論・質疑を踏まえ、自分たちの優位性を理由や具体例を添えて分かりやすく説明している。
	8	立論・質疑を踏まえ、自分たちの優位性を改めて説明している。
	4	立論・質疑を踏まえ、自分たちの優位性を改めて説明しようとしているが、内容が不明瞭である。
12点	We believe that high school students are mature enough to handle both a job and their studies, so there's no need to prohibit them from working part time. Rather, through keeping a good balance between study and work, they can learn how to manage time.	
8点	We believe that high school students can keep a good balance between study and work. They are high school students, so they know what they have to do. No problem.（具体性に欠ける）	

【総括】(Summary)

内容	12	相手の主張と比較し、自分たちの主張の方が重要性が大きいことを、具体的な根拠とともに分かりやすく説明している。
	8	相手の主張と比較し、自分たちの主張の方が重要性が大きいことを説明している。
	4	相手の主張と比較し、自分たちの主張の方が重要性が大きいことを印象付けようとしているが、内容が不明瞭である。
12点	Of course, working part time has some risks like affecting their academic performance, but, in a democratic society, everyone has the right to do what they like, within the law. If students say that they can balance their schoolwork with part-time jobs, no one can stop them from doing what they want. The experience of balancing the two will help them become responsible and independent adult.	
8点	Certainly, there are some risks about working part time, but if students say they want to work part time, who can stop them? We need to believe they will do well both in study and in work. The experience of part-time job will bring them a lot of benefits.（抽象的）	

付表C

ディスカッションのトピック

1. Why do you think some students study abroad?
2. Which do you prefer to work for, a large company or a small company?
3. Do you think games are important for adults as well as for children?

4. Why do you think some people are attracted to dangerous sports or activities?
5. Which do you think is better, being single or being married?
6. Which do you think is better, working by hand or using machines?
7. What is the most important factor to be successful in life?
8. What would you like to be if you were born again?
9. Which is more important to learn, history or science?
10. Where is the most important place in your house?
11. Do you agree with the following statement? Everyone must go to university.

付表D

ディベートのトピック

1. Students should be asked to study foreign languages other than English.
2. There should be boys-only and girls-only high schools in addition to co-education schools.
3. We should abolish smoking in all restaurants.
4. We should ban Giri chocolate.
5. High school shouldn't allow students to bring their cell phones to school.
6. High achieving students should be allowed to skip grades.
7. Students should be allowed to choose their homeroom teacher.
8. All students should join a club.
9. There should be a convenience store in high schools.
10. Animal testing should be banned.
11. Zoos should be abolished.
12. There should be no homework for high school students.

付表E

ディスカッションの場合のMFRMのインプットファイル　　注. ( ) = 解説

title = Discussion_rater_criteria
convergence = 0.1 ; size of largest remaining marginal score residual at convergence
unexpected = 2 ; size of smallest standardized residual to report
arrange = m ; arrange output tables in Num decending and Logit ascending order

facets = 3 ; 3 facets 1 Person, 2 Rater, 3 Criteria
noncenter = 1 ; examinee facet floats
positive = 1 ; for examinees, greater score greater measure
Pt-biserial = Yes ; report the point-biserial correlation
Inter-rater = 2 ; facet 2 is the rater facet
Missing = N
Yardstick = 0,2,-5,5
Model =
?,?,1,R3
?,?,2,R3
?,?,3,R3
*

Labels=
1,Ss
001-227

*
2,Rater
01 = Rater1
02 = Rater2
03 = Rater3

*
3,Criteria
1 = Content
2 = Expression
3 = Technique

*
data =
001 01 1-3 3 3 3
（受験者番号001、採点者01、観点3つ1-3、観点ごとのスコア3つの順）
002 01 1-3 3 2 2
（途中略）

001 02 1-3 3 2 2
（途中略）
001 03 1-3 N N N
227 03 1-3 2 2 1

（最後にエンターキーを入れる）

付表F

ディスカッションの場合のmG theoryのインプットファイル

GSTUDY p x r Design with Covariance Components Design = p
OPTIONS "*.out"
MULT 3 Con Expr Techni
EFFECT * p 225 225 225
（227名中2名は採点者1名のみの採点だったため225名で分析）
EFFECT # r 2 2 2
FORMAT 0 0
PROCESS
3 3 3 2 3 2
（Contentの採点者1、2のスコア、Expressionの採点者1、2のスコア、Techniqueの採点者1、2のスコアの順）
（略）
DSTUDY p x R Design with Covariance Components Design = p
DOPTIONS DCUT 2.0
DEFFECT $ p 225 225 225
DEFFECT # R 3 3 3
ENDDSTUDY
DSTUDY p x R Design with Covariance Components Design = p
DOPTIONS DCUT 2.0
DEFFECT $ p 225 225 225
DEFFECT # R 2 2 2
ENDDSTUDY
DSTUDY p x R Design with Covariance Components Design = p
DOPTIONS DCUT 2.0
DEFFECT $ p 225 225 225
DEFFECT # R 1 1 1

ENDDSTUDY

（最後にエンターキーを入れる）

付表G

ルーブリックの適切さの判断

表G1には研究全体の結果をまとめた。

ルーブリックの適切さの判断基準は、5点ある（表G1参照）。その基準を用いた解釈例として、「ディスカッションの内容」の結果を以下に挙げる（図G1と図G3左参照）。また図G2と図G3右に、問題が2点見つかった「ディベートの音量等」の結果を示す。

表G1.
ルーブリックの適切さの結果

基準	各レベル難易度・敷居推定値	各レベル使用頻度	各レベルの適合度	隣接する敷居推定値の距離		ルーブリックの確率曲線
	段階的に上昇	10以上	Outfit MS 2.0未満	1.4以上5.0以内（括弧内は値）		各レベルに頂上あり
ディスカッション						
内容	OK	OK	OK	OK	(4.66)	OK
表現	OK	OK	OK	問題あり	(5.08)	OK
技術点	OK	OK	OK	問題あり	(5.94)	OK
ディベート（内容個人点）						
内容	OK	OK	OK	OK	(4.66)	OK
姿勢	OK	OK	OK	OK	(5.00)	OK
文法・語法	OK	問題あり [4]	OK	OK	(4.88)	OK
音量等	OK	問題あり [2]	OK	問題あり	(7.12)	OK
ディベート（内容グループ点）						
内容	OK	問題あり [0]	OK	--		--
姿勢	OK	OK	OK	OK	(4.42)	OK
文法・語法	OK	問題あり [4]	OK	OK	(4.28)	OK
音量等	OK	問題あり [2]	OK	問題あり	(6.46)	OK

注. 問題あり = レベル 1 で問題あり。[ ] = レベル1で観測された頻度。-- = レベル1のスコアがなかったため、算出・描画されなかった。

基準1:各レベルの難易度推定値と敷居推定値が段階的に上昇するか

「ディスカッションの内容」について、図G1では、Average measures (Avge Meas)が-2.06から3.34まで上昇し、RASCH-ANDRICH Threshold Measureも-2.33から2.33まで上昇し、ともに段階的に上昇していた。

基準2:各レベルの使用頻度は、10回以上あるか

各レベル使用頻度は、Counts Usedが26, 163, 175と各レベルで10回以上あった。

基準3:各レベルの適合度は、アウトフィット平均平方(Outfit mean squares)が2.0未満か

OUTFIT MnSqで0.9～1.1で見たしていた。

基準4:隣接する敷居値の距離は1.40以上、5.00以内か

-2.33と2.33の値から距離は4.66と算出でき、基準は満たしていた。

基準5:ルーブリックの確率曲線(図G1)においては各レベルに頂上が見えるか

レベル2に頂上が見えていた。

図G1.

「ディスカッションの内容」の「ルーブリックの適切さ」の結果

DATA				QUALITY CONTROL		RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category	Counts		Cum.	Avge	Exp.	OUTFIT	Thresholds	Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	28	26	7%	-2.06	-1.97	.9			(-3.40)		low	low	100%
2	163	163	45%	1.20	1.14	1.0	-2.33	.26	.00	-2.33	-2.33	-2.33	84%
3	261	175	48%	3.34	3.39	1.1	2.33	.14	(3.41)	2.35	2.33	2.32	100%
			100%						(Mean)		(Modal)	(Median)	

図G2.

「ディベートの音量等」の「ルーブリックの適切さ」の結果

DATA				QUALITY CONTROL		RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category	Counts		Cum.	Avge	Exp.	OUTFIT	Thresholds	Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	2	2	1%	2.19	-.06	1.4			(-4.62)		low	low	100%
2	216	216	55%	2.53	2.42	1.2	-3.56	.72	.00	-3.54	-3.56	-3.56	95%
3	233	174	44%	3.86	4.02	1.2	3.56	.12	(4.64)	3.55	3.56	3.55	100%
			100%						(Mean)		(Modal)	(Median)	

図G3.
「ディスカッションの内容」（左）と「ディベートの音量等」（右）の確率曲線