

日本人英語学習者を対象としたタスクに基づくライティング評価法の開発

*The Development of a Construct-based
Processing Approach to Testing:
Task-based Writing Assessment
for Japanese Learners of English*

山梨県立大学 杉田由仁

第1章 序論

第1節 研究の背景

- 1) 1970年代からのL2ライティング指導の変遷
- 2) 日本の学校におけるライティング指導の変容
- 3) 『学習指導要領』の改訂

第1章 序論

第2節 研究主題(Research Question)

1. What is the definition of the writing abilities to be assessed?
2. How are writing tasks designed?
3. What types of procedure are more suitable to construct rating scales?
4. To what extent, statistically, is the task-based writing test a reliable and valid measure?

第2章 ライティング評価法の背景

第1節 はじめに

本章では、タスクに基づくライティングテスト(TBWT)開発の基礎となるライティング評価法に関わる先行研究の概観とテスト開発の理論的枠組みおよびライティング能力を構成する概念の検討を行っている。

第2章 ライティング評価法の背景

第2節 Canale & Swain Approach

「文法能力」「談話能力」「社会言語能力」「方略的能力」が言語能力テスト開発の基本的枠組みとなり得ることを示すことはできたが、テストパフォーマンス時における測定方法を提示することができていない。

第2章 ライティング評価法の背景

第3節 Bachman model

「方略的能力」の再定義により、言語使用場面における言語能力の機能を明確化した。テスト得点は実際の言語使用場面における言語能力を示す指標と解釈することが可能となる。しかし、言語処理における認知的側面への配慮に課題が残る。

第2章 ライティング評価法の背景

第4節 Skehan's perspective

第二言語知識体系は、規則に基づく体系と記憶に基づく体系(二層性の体系)を持ち、両者はトレードオフの関係にある。評価タスクの困難度を調節することにより、言語運用能力を測定することが可能となる。

第2章 ライティング評価法の背景

第5節 Bachman's claims

タスクの困難度とは、パフォーマンスにおけるさまざまな要因と関連するので、その調節により評価タスクのレベル設定を行うという方法は現実的ではない。測定対象となる構成概念を定義することが必要である。

第2章 ライティング評価法の背景

第6節 Construct-based processing approach (TBWT開発の枠組み)

- 1) 測定対象となる構成概念の明確化
- 2) 言語処理要因を考慮したタスクデザインの検討
- 3) 1) 2) を包括する評価タスクの開発と予備調査の実施

第2章 ライティング評価法の背景

- 1) 規則に基づく体系への依存度が高い“accuracy,” 記憶に基づく体系への依存度が高い“communicability”を、「書く」領域における言語運用能力の構成概念とする
- 2) input, expected response の調節により、評価タスクの特徴づけを行う

第2章 ライティング評価法の背景

第7節 TBWTの有用性検証

本研究では、古典的信頼性・多相ラッシュ測定に基づく信頼性および構成概念妥当性により、TBWTの有用性について検証を行う。

第3章 評価タスクと評定尺度の開発

第1節 はじめに

本章では、評価タスクおよび評定尺度開発の具体的手順を3段階に分けて説明する。

第3章 評価タスクと評定尺度の開発

第2節 タスクのデザインと特徴づけ

Task 1 (accuracy): content-based support, form-focused stakes

→ writing a letter of self-introduction

Task 2 (communicability): form-oriented support, meaning-focused stakes → discussion task

第3章 評価タスクと評定尺度の開発

第3節 主なライティング評価法の検討

全体的評価法: TWE, FCE

分析的評価法: ESL Profile

単特性評価法: NAEP Scoring Guide

多特性評価法: Michigan Writing Assessment

第3章 評価タスクと評定尺度の開発

第4節 評定尺度の作成

“accuracy” “communicability”それぞれの定義に適合すると考えられる記述子を主要な尺度から抽出し、構成概念ごとに分類・整理統合して、2つの6段階評定尺度を作成する。

第4章 予備テストの結果と分析

本章では、開発された評価タスクおよび評定尺度を用いて実施された予備調査における5名の現職高校教師による評定とその分析結果について詳述する。

研究目的

タスクに基づくライティングテストに関連する要因(評定者、被験者能力、タスク)が、相互に影響し合う度合いについて分析を行う。

評価タスクおよび評定尺度の信頼性について検証する。

研究方法

評価対象データの収集

対象:A大学教育人間科学部1年生15名
2種類のタスク

「自己紹介の手紙(100~120語)」 20分

「ディスカッションのためのメモ」 10分

辞書の使用は認めない

ライティング・タスク 1

You are going to stay with Parker Family in Britain this summer.
Write a 100-120 word letter introducing yourself to your host family. Before writing, think of the following topics.

- Your name and age
- Your job, major in school

ライティング・タスク 1

- Your family and pets
- Your interests and hobbies
- Your favorite places, foods, activities
- Your experience in traveling abroad
- Some things you want to do while you are in Britain

タスク1 (Accuracy)の評定尺度

Accuracy: A (6)

- The written text
- is well organized and well developed (TWE).
 - shows strong rhetorical control and is well managed (MWA).
 - has clear organization with a variety of linking devices (FCE).

タスク 1 (Accuracy)の評定尺度

Accuracy: A (6)

- The written text
- demonstrates appropriate word choice though it may have occasional errors (TWE).
 - has few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions, spelling, punctuation, capitalization, paragraphing (ESL).

ライティング・タスク 2

You will have 10 minutes to make notes about the following discussion topic, "Why do you study English?" In order to prepare for the discussion, think of answers to the question as many as possible and write them as "To travel abroad."

タスク 2 (Communicability)の評定尺度

Communicability: A (6)

- The written text
- displays consistent facility in use of language (TWE).
 - contains well-chosen vocabulary to express the ideas and to carry out the intentions (MWA).

タスク 2 (Communicability)の 評定尺度

Communicability: A (6)

The written text

-effectively addresses the writing task (TWE).

-has a very positive effect on the target reader with adequately organized relevant ideas (FCE).

研究方法

分析データの収集

対象: 関東近県の公立高校において英語科を担当する5名の現職教師

依頼した評定作業の内容:

タスク1 (accuracy) の評定

タスク2 (communicability) の評定

全体的印象による総合評価

分析方法

FACETS プログラムによって検出される標準誤差(standard error)とフィット統計(fit statistics)により、「評定者」「被験者能力」「タスク」という3つの要因(facets)が予測モデルに適合しているか確認する

分離指数(separation index)および信頼性係数(reliability coefficient)に基づいて、評価タスクと評定尺度の信頼性を検討する

結果: 尺度の一貫性

評定尺度の一貫性

タスク1評価基準による評定: $\alpha = 0.9349$

タスク2評価基準による評定: $\alpha = 0.9466$

全体的印象による総合評価: $\alpha = 0.9443$

→ α 信頼性係数が、妥当な評定尺度とみなすための規準, 0.8以上 (Davies, 1990)となり、評定尺度としての内的一貫性が確認された。

結果: 評定者について

評定者	1	2	3	4	5
厳しさ	0.35	1.03	0.05	-0.03	-0.76
誤差	0.21	0.21	0.21	0.21	0.22
Infit	0.94	0.73	0.83	0.77	1.44

→ 評価者は全員が一貫した評定を行い、評定が一致した割合(34.5%)は、各評定者が "independent experts" として評価できたことを意味する

結果: タスクの難易度

タスク	1	2	印象
難易度	-0.18	0.14	0.40
誤差	0.16	0.16	0.17
Infit	1.01	1.10	0.68

→ ロジット得点による難易度: $\chi^2(2)=2.0, p=.36$
タスクの評定値および全体的評価に、難易度の違いによる影響は認められない

結果:タスクの信頼性(accuracy)

評定者	測定値	期待値	差	バイアス	Zスコア
1	47	48.4	-.10	-.18	-.49
2	55	53.9	.08	.13	.37
3	50	46.0	.28	.50	1.43
4	41	45.4	-.31	-.61	-1.58
5	41	40.2	-.05	-.11	.29

→Zスコアは「有意でないバイアス(-2<Z<2)」の範囲内にあり、評定者は独自のバイアスを持たずに評定できた。

結果:タスクの信頼性(communicability)

評定者	測定値	期待値	差	バイアス	Zスコア
1	52	51.9	.01	.01	.04
2	56	57.1	-.08	-.15	-.40
3	46	49.6	-.26	-.46	-1.29
4	52	48.9	.22	.40	1.11
5	45	43.3	.12	.22	.60

→Zスコアは「有意でないバイアス(-2<Z<2)」の範囲内にあり、評定者は独自のバイアスを持たずに評定できた。

結果:全体評価の信頼性(Impression)

評定者	測定値	期待値	差	バイアス	Zスコア
1	52	50.6	.10	.19	.52
2	56	55.7	.02	.04	.12
3	46	48.3	-.02	-.05	-.12
4	52	47.7	.09	.18	.48
5	45	42.6	-.18	-.39	-.98

→Zスコアは「有意でないバイアス(-2<Z<2)」の範囲内にあり、評定者は独自のバイアスを持たずに評定できた。

結果:評定尺度(Accuracy)の信頼性

評定	度数	%	STEP	Outfit
1	4	6		.4
2	15	21	-4.48	.9
3	22	31	-1.36	1.0
4	16	23	.83	1.2
5	8	11	2.26	.6
6	5	7	2.74	1.4

Step difficulty の差: 5→6 (=0.48) < 1.4 については、数値基準が満たされていない

結果:評定尺度(Communicability)の信頼性

評定	度数	%	STEP	Outfit
1	3	4		1.9
2	17	24	-4.74	.8
3	14	20	-.66	1.3
4	13	19	.52	.8
5	18	26	1.22	.9
6	5	7	3.66	1.5

Step difficulty の差: 3→4 (=1.18), 4→5(=0.7) < 1.4 については、数値基準が満たされていない

結果:評定尺度(Impression)の信頼性

評定	度数	%	STEP	Outfit
1	3	4		.7
2	15	21	-4.81	.9
3	17	24	-1.11	.6
4	18	26	.39	.4
5	13	19	1.93	.7
6	4	6	3.59	.8

Outfit数値、Step difficulty の差のすべてについて、数値基準が満たされている

まとめと課題

各評定者はタスクに対して独自の評価バイアスを持たずに一貫した評定を行うことができたことがわかった。

それぞれの評定尺度には一定の信頼性があることが確認されたが、尺度の等間隔性には課題が残され、6段階から5段階(または4段階)の尺度に改訂する必要がある。

第5章 評定尺度の改訂

第1節 はじめに

本章では、予備テストの分析結果に基づき、評定尺度の改訂作業を行う。

(本章の一部は、刊行論文: Developing and improving rating scales for a task-based writing performance test, *JLTA Journal* No. 12, pp.85-104 と重複する)

第5章 評定尺度の改訂

第2節 改訂作業の手順

Step 1: accuracy のカテゴリ-5・6を結合する

Step 2: communicability については3・4(または4・5)を結合する

Step 3: FACETS 分析を行う

Step 4: ガイドラインの基準値と照合する

第5章 評定尺度の改訂

第3節 各モデル尺度の適合度

カテゴリ-5・6を結合したaccuracyの5段階尺度はガイドラインに示された一部の基準と適合しない面があったが、3・4を結合したcommunicabilityの5段階尺度はすべての基準に適合した。

第5章 評定尺度の改訂

第4節 「評定の手引き」の修正作業

1) 評定尺度の精度をより向上させるために、accuracy, communicability の下位概念を Organizational skills と Linguistic accuracy, Communicative quality と effect にそれぞれ細分化した。

タスク1の構成概念

Accuracy	
Organizational skills	Linguistic accuracy
The writing displays a logical organizational structure which enables the content to be accurately grasped.	Errors of vocabulary, spelling, punctuation or grammar.

タスク1 (Accuracy)の評定尺度

Organizational skills

The written text

- is well organized and well developed (TWE).
- shows strong rhetorical control and is well managed (MWA).
- has clear organization with a variety of linking devices (FCE).

タスク 1 (Accuracy)の評定尺度

Linguistic accuracy

The written text

- demonstrates appropriate word choice though it may have occasional errors (TWE).
- has few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions, spelling, punctuation, capitalization, paragraphing (ESL).

タスク2の構成概念

Communicability

Communicative quality	Communicative effect
The writing displays an ability to communicate without causing the reader any difficulties.	Quantity of ideas to develop the response and relevance of the content to the proposed task.

タスク 2 (Communicability)の評定尺度

Communicative quality

The written text

- displays consistent facility in use of language (TWE).
- contains well-chosen vocabulary to express the ideas and to carry out the intentions (MWA).

タスク 2 (Communicability)の評定尺度

Communicative effect

The written text

- effectively addresses the writing task (TWE).
- has a very positive effect on the target reader with adequately organized relevant ideas (FCE).

第5章 評定尺度の改訂

第4節 「評定の手引き」の修正作業

- 3) 予備調査のサンプルを評価基準に「きわめてあてはまる(5)」から「あてはまらない(1)」の5段階に再分類し、各段階を代表するサンプルを選定する
- 4) 選定したサンプルに評価基準を具体的に説明する「解説」をつける

第5章 評定尺度の改訂

第5節 作業結果のまとめ

評定尺度改訂の作業結果をまとめ、5段階評定尺度ならびに評定の手引き(改訂版)の信頼性・妥当性検証の必要性について強調した。

第6章 本調査(1)の結果と分析

本章では、予備調査と同一の5名の現職高校教師による本調査(1)の結果とその分析について詳述する。
(本章は修正の上、The Development and implementation of task-based writing performance assessment, PAAL Vol. 13(2), pp.85-104 として刊行している)

研究目的

タスクに基づくライティングテスト(TBWT)に関連する要因(評定者、被験者能力、タスク)が、相互に影響し合う度合いについて分析を行う。
改訂されたTBWTの信頼性・妥当性について検証する。

研究方法

評価対象データの収集

対象:Y県立大学国際政策学部20名(2年生14名、3年生6名)

2種類のタスク

「自己紹介の手紙(100~120語)」20分
「ディスカッションのためのメモ」10分

Criterionによるエッセイ

“Why do you think people attend college or university?” 30分

研究方法

分析データの収集

対象:関東近県の公立高校において英語科を担当する5名の現職教師(予備調査と同一の評定者)

依頼した評定作業の内容:

タスク1 (accuracy) の評定

タスク2 (communicability) の評定

全体的印象による総合評価

分析方法

FACETS プログラムによって検出される標準誤差(standard error)とフィット統計(fit statistics)により、「評定者」「被験者能力」「タスク」という3つの要因(facets)が予測モデルに適合しているか確認する
分離指数(separation index)およびRater × Subjects, Rater × Tasks のバイアス分析に基づいて、評価タスクと評定尺度の信頼性を検討する

結果:尺度の一貫性

評定尺度の一貫性

タスク1評価基準による評定: $\alpha = 0.9386$

タスク2評価基準による評定: $\alpha = 0.9582$

全体的印象による総合評価: $\alpha = 0.9570$

→ α 信頼性係数が、妥当な評定尺度とみなすための規準, 0.8以上 (Davies, 1990)となり、評定尺度としての内的一貫性が確認された。

結果:評定者について

評定者	1	2	3	4	5
厳しさ	0.35	0.52	0.97	-0.69	-1.41
誤差	0.24	0.24	0.24	0.24	0.25
Infit	0.97	0.93	0.78	0.65	1.16

→分離指数(0.92), χ^2 検定の結果、5名の評定者の厳しさは同等ではなかったが、Infit 数値は0.90 ± [0.17 × 2]の範囲にあることから、全員評定者内一貫性があった

結果:評定者1のバイアス分析

被験者	測定値	期待値	差	バイアス	Zスコア
12 (U)	9	13.2	-1.41	-4.26	-3.72
19 (M)	12	9.7	0.77	2.25	2.35

→評定者1は「語数の超過」「文章構成力の不足」があるとaccuracy の評定に「厳しい」傾向が、「言語的正確さ」があると「あまくなる」傾向がある。Communicability については「回答項目が類似して項目数が少ない」と「厳しい」傾向が、「情報伝達の効果」があると「あまくなる」傾向がある。

結果:評定者5のバイアス分析

被験者	測定値	期待値	差	バイアス	Zスコア
12 (U)	14	11.3	0.89	2.65	2.27
7 (M)	8	10.0	-0.65	-2.40	-2.24

→評定者5は「語数が多く」「文章構成力」があるとaccuracy の評定に「あまくなる」傾向が、「文章構成力の不足」があると「厳しい」傾向がある。Communicability については「回答項目が類似して項目数が少ない」と「厳しい」傾向が、「情報伝達の効果」があると「あまくなる」傾向がある。

結果:評定者3のバイアス分析

被験者	測定値	期待値	差	バイアス	Zスコア
4 (L)	7	5.0	0.68	2.53	2.38

→評定者3は「文章構成力の不足」があるとaccuracy の評定に「厳しい」傾向が、また Communicability の下位能力である「情報伝達の効果」があると「あまくなる」傾向がある。

結果:タスクの難易度

タスク	1	2	印象
難易度	0.13	-0.18	0.50
誤差	0.19	0.19	0.19
Infit	1.10	0.92	0.68
弁別推定値	0.90	1.05	1.37

→ロジット得点による難易度: $\chi^2(2)=1.5, p=.47$
 タスクの評定値および全体的評価に、難易度の違いによる影響は認められない
 →弁別推定値(Estimate of Discrimination)
 $0.5 < E.D. < 1.5$ モデルに適合

結果:タスクの信頼性(accuracy)

評定者	測定値	期待値	差	バイアス	Zスコア
1	64	62.4	.08	.28	.66
2	62	63.4	-.07	-.24	-.58
3	70	66.0	.20	.72	1.69
4	57	56.5	.02	.08	.20
5	48	52.6	-.23	-.90	-2.01

→評定者1~4のZスコアは「有意でないバイアス(-2<Z<2)」の範囲内にあり、独自のバイアスを持たずに評定できた。

結果:タスクの信頼性(communicability)

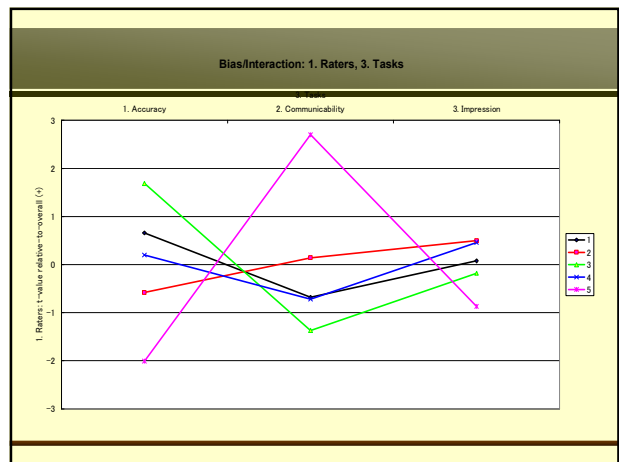
評定者	測定値	期待値	差	バイアス	Zスコア
1	59	60.7	-.08	.28	-.68
2	62	61.7	.02	.06	.14
3	61	64.3	-.17	-.56	-1.37
4	53	54.7	-.08	-.31	-.72
5	57	50.7	.31	1.13	2.70

→評定者1~4のZスコアは「有意でないバイアス(-2<Z<2)」の範囲内にあり、独自のバイアスを持たずに評定できた。

結果:全体評価の信頼性(Impression)

評定者	測定値	期待値	差	バイアス	Zスコア
1	62	61.8	.01	.03	.08
2	64	62.8	.06	.20	.50
3	65	65.4	-.02	-.07	-.18
4	57	55.9	.05	.19	.46
5	50	52.0	-.10	-.38	-.87

→Zスコアは「有意でないバイアス(-2<Z<2)」の範囲内にあり、5名の評定者全員が独自のバイアスを持たずに評定できた。



評価タスクの妥当性

	R1	R2	R3	R4	R5	Av.
Task 1	.71	.70	.65	.63	.60	.66
Task 2	.74	.71	.67	.70	.79	.72
Impression	.74	.78	.70	.72	.68	.72

各評定者のタスクごとの得点とCriterionによる6段階の得点との相関係数 → 一定程度の基準関連妥当性が認められる

結果:評定尺度(Accuracy)の信頼性

評定	度数	%	STEP	Outfit
1	10	10		.9
2	23	23	-5.28	1.1
3	33	33	-1.54	1.4
4	24	24	1.96	.8
5	10	10	4.86	1.0

Step difficulty の差 (1.4 logits 以上 5.0 logits 以下)
: 数値基準が満たされている
Outfit 値 (2.0 以下): カテゴリーがモデルに適合

結果: 評定尺度(Communicability)の信頼性

評定	度数	%	STEP	Outfit
1	13	13		1.2
2	23	23	-4.97	1.6
3	33	33	-1.49	.9
4	21	21	1.98	.6
5	10	10	4.49	.7

Step difficulty の差 (1.4 logits 以上 5.0 logits 以下)
: 数値基準が満たされている
Outfit 値 (2.0 以下): カテゴリーがモデルに適合

結果: 評定尺度(Impression)の信頼性

評定	度数	%	STEP	Outfit
1	11	11		.7
2	22	22	-5.13	.5
3	36	36	-1.65	.7
4	20	20	2.27	.7
5	11	11	4.51	.8

Step difficulty の差 (1.4 logits 以上 5.0 logits 以下)
: 数値基準が満たされている
Outfit 値 (2.0 以下): カテゴリーがモデルに適合

まとめ (テストの信頼性・妥当性)

各評定者は一貫性ある評定を行ったが、厳しさは同等ではなかった。
それぞれの評定尺度には一定の信頼性があり、これら尺度による評定結果には、評価タスクの難易における有意差は観察されなかった。
評価タスクはモデルに適合し、弁別妥当性が認められる。
Criterion と評価タスクによる評定値の相関から、基準関連妥当性が認められる。

今後の課題

評定者1・3・5は特定の被験者に対して独自のバイアスを持って評定を行った。
評定者5は、タスク1・2に対して独自の評価バイアスを持って評定を行った。
バイアス分析の結果について検証を行うために、異なる評定者グループを対象として、比較調査を行う必要がある。

第7章 本調査(2)の結果と分析

本章では、5名の現職中学校教師による本調査(2)の結果とその分析についての説明を行う。

研究目的

タスクに基づくライティングテスト(TBWT)に関連する要因(評定者、被験者能力、タスク)が、相互に影響し合う度合いについて分析を行う。
予備調査・本調査(1)とは異なる評定者グループを対象として、改訂されたTBWTの信頼性・妥当性について検証する。

研究方法

評価対象: 本調査 (1) と同じデータ
対象: Y県立大学国際政策学部20名(2年生14名、3年生6名)
2種類のタスク
「自己紹介の手紙(100~120語)」 20分
「ディスカッションのためのメモ」 10分
Criterion によるエッセイ
“Why do you think people attend college or university?” 30分

研究方法

分析データの収集
対象: 山梨県内の公立中学校において英語科を担当する5名の現職教師
依頼した評定作業の内容:
タスク1 (accuracy) の評定
タスク2 (communicability) の評定
全体的印象による総合評価

結果

各評定者はタスクに対して一貫した評定を行うことができた。
各タスクに対応する5段階評定尺度には一定の信頼性があることが確認された。
5名の評定者全員に、特定の受検者に対する特有のバイアスが認められた。
5名の内1名には、特定のタスクおよび特定の受検者に対するバイアスが認められた。

まとめと課題

異なる評定者グループに対しても、TBWTには一定の信頼性・妥当性があることが確認された。
TBWT に対する熟練度の違いという観点から、本調査 (1) (2) の10名の評定者について比較分析を行う必要がある。

第8章 比較調査の結果と分析

本章では、本調査 (1) (2) の各5名の高校教師および中学校教師による評定結果について、比較分析を行った結果について記述している。

研究目的

タスクに基づくライティングテスト (TBWT) に対する熟練度が、評定の厳しさ、評定の一貫性に影響する度合いについて分析を行う。
評定者の熟練度と被験者能力、タスクが、相互に影響し合う度合いについてバイアス分析を行う。

研究方法

分析データ

本調査 (1) および (2) において、各5名の中学・高校教師が行った
「タスク1 (accuracy) の評定」
「タスク2 (communicability) の評定」
「全体的印象による総合評価」

結果

熟練度の差によらず各評定者はタスクに対して一貫した評定を行うことができた。熟練度の高い高校教師の方が評定があまり傾向が確認められた。
熟練度の低い中学校教師の方が一貫性に欠ける面が見られたが、グループ間の差には至らなかった。
バイアス分析においては、特定のタスクとの相互作用に若干の差が認められた。

まとめ

TBWT に対する熟練度の違いは、評定の厳しさについては差違を生じさせる影響を与えたが、評定の一貫性については差違を生じさせる影響を与えなかった。
「評定の手引き」は、熟練度の差によらず各評定者が一貫した評定を行う上で、一定の効果があった。

第9章 考察

本章では、全体の主題として提示した4つのリサーチ・クエスチョンを議論の柱として、調査結果をもとに総合的な考察を行う。

RQ. 1 ライティング能力の定義

タスクに基づくライティングテスト (TBWT) の評価対象となる言語能力特性は、accuracy, communicability である。
TBWT を日本の教室に導入することにより、学校における英語科指導法やカリキュラムに波及効果が与えられる可能性がある。

RQ. 2 評価タスクのデザイン方法

TBWTにおける評価タスクのデザイン方法は、測定対象となる構成概念の定義と言語処理の側面を考慮した“Construct-based processing approach”である。
評価タスクに特徴づけを行い、言語処理に関わる要因を操作することにより目標とする言語能力特性を測定する。

RQ. 3 評価尺度の作成方法

TBWT の評価尺度は、accuracy と communicability という多特性に対してそれぞれ評価基準を設定するという単特性評価法の手法を取り入れて作成が行われる。

評定者は、1人の受検者に対して2課題の評定を行う必要があるが、分析的評価法ほど時間がかからず、信頼性のある評定を行うことができる。

RQ. 4 タスクに基づく評価の信頼性・妥当性

TBWT の評価タスクおよび評価尺度によるライティング能力評価には、一定の信頼性がある。

評価タスクはモデルに適合し、弁別妥当性が認められ、また *Criterion* と評価タスクによる評定値の相関から、基準関連妥当性が認められる。さらに、評定の手引きに対する評価結果などから、一定の妥当性が確保されていると考えられる。

評定者トレーニングのあり方

TBWT の信頼性および妥当性をより向上させるための具体的方法として評定者トレーニングをどのような内容・方法で行うことが望ましいか、について考察を行っている

今後のTBWT開発に向けての提言

1. 学習指導要領に基づいて評価タスクをより多く案出すること
2. TBWT 導入に伴う波及効果を検証すること
3. Construct-based processing approach によるテスト開発を継続すること
4. 実用性について検証すること
5. 評価尺度および評定の手引きの改訂
6. 大規模調査および評定者の認知過程に対する洞察を行うこと

第10章 結論

本研究の成果

- 1) 評価対象となるライティング能力の明確な定義
- 2) 日本人学習者を対象としたタスクに基づくライティングテスト開発方法の確立
- 3) 信頼性・妥当性の検証による有用性の高いパフォーマンステストの提供

第10章 結論

教育的示唆

- 1) TBWTの内容構成と同様に、accuracyから communicability (fluency)へ向かうライティング指導が行われることが望ましい。
- 2) ライティングのパフォーマンステストを積極的に教室での指導に導入することが望ましい。
- 3) 教師教育の一環として、ライティングの評定者トレーニングを取り入れることが望ましい。

第10章 結論

本研究の限界

- 1) ライティングのサンプル数、評定者の人数ともに不十分である。大規模調査による再検証が必要である。
- 2) 評価タスクをより多く案出し、タスクと評定尺度を個別の facets として扱い、それぞれがスコアに与える影響を分析する必要がある。
- 3) 評定セッションを対面式で行い、評定者の認知過程に関する調査を行う必要がある。

第10章 結論

今後の課題

- 1) 中等教育段階における大規模調査の実施
- 2) 評定者の評定作業における認知過程に対する調査の実施
- 3) テスト法としての有用性に関する調査の実施