

タスクに基づくライティングテストによる 言語運用能力の評価と測定

ー評価タスクと評定尺度の信頼性の分析ー

山梨県立大学 杉田由仁

はじめに

- 第二言語知識体系:規則に基づく体系と記憶に基づく体系(二層性の体系)、トレードオフの関係(Shehan 2001)
- 「書く」領域における言語運用能力の構成概念: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”
- 評価タスク・評定尺度の開発と予備調査

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.

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.

研究方法

- 評価尺度の作成
 - 1) TOEFL, FCE, Michigan の評価基準を accuracy, communicability の概念規定と照合する
 - 2) 対応する評価基準を選定し、分類・整理統合して6段階の評価尺度を作る

タスク1 (Accuracy) の評価尺度

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) の評価尺度

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 (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).

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、評定が一貫した割合 (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

まとめと今後の課題

- 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。