第二言語読解能力評価におけるリーダビリティの活用法

小谷 克則† 吉見 毅彦‡/† 九津見 毅†† 佐田 いち子†† 井佐原 均†
†情報通信研究機構 ‡龍谷大学 ††シャープ株式会社
kat@khn.nict.go.jp

1. はじめに

近年，インターネットの普及などにより，英文テキストを読む機会が急速に増えつつある．また，大学英語教育において，各専門分野に特化した論文読解などを中心とした授業が増えていることからも，これまでの英語の教科書を読むだけでなく，論文などの実際的な文章をある程度の量読むことが必要とされているといえる．

実践的な英文読解を行う場合，内容理解の度合いはもちろん，読解の速さも重要となる．したがって，実践的な英文読解能力を養成することを目標として学習者の読解能力を評価する場合，理解度と読解速度の積である読解効率を評価指標とした読解効率テストを利用する必要がある[1]，[2]．

実践的な英文読解を目標とする授業において読解効率テストを行う場合，インターネット上のテキストや論文のように多種多様のテキストがテストの対象となる．これらのテキストを読解効率テストに利用する場合，テキストの読みやすさが異なることが一つの問題点として考えられる．読解効率テストの評価指標の一つである読解速度はテキストの読みやすさの影響を強く受けると考えられている[3]．そのため，読みやすさが異なるテキストが混在するテストと読みやすさが統一されているテキストにより構成されるテストとでは，テストの有効性が異なると考えられる．

そこで，本稿は読解効率テストの有効性がテストに含まれるテキストにより異なるのかどうかを確かめるために実験を行った．本実験において，次の三種類のテキスト群から構成される読解効率テストの有効性を検証した．同一テキストにより構成される読解効率テスト（同一テスト），同程度の読みやすさの異なるテキストにより構成される読解効率テスト（同程度テスト），そして，読みやすさもテキストも異なるテキストから成る読解効率テスト（異なりテスト）である．実験の結果，読みやすさによる影響よりも事例数による影響が大きいことがわかった．

2. テストの有効性

古典的テスト理論では，テストの有効性はテストの信頼性と妥当性を検証することによって確認できると考えられている[4]．信頼性の高いテストとは，能力が同程度であるテスト受験者に同じテストを受けさせた時の結果が近ければ近いほど高くなると考えられている．本実験では，読解効率テストの信頼性をクロンバック α 信頼性係数により検証した．クロンバック α 信頼性係数とは，同一テストによる検証法とは異なり，一つのテストを分割した場合の信頼性を示す値である[1]．

クロンバック α 信頼性係数は式（1）により算出される．式（1）において k はテストの項目数であるテキスト数，S_j は各テキストにおける読解効率の標準偏差，S_Y は全テキストにおける読解効率の標準偏差である．

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} S_j^2}{S_Y^2}\right) \qquad (1)$$

テストの妥当性は，テスト受験者の能力の評価がより適正であればあるほど高くなると考えられている（目標関連妥当性）[2]．この目標関連妥当性は，既存の有効性が確認されているテストの結果と対象となるテスト結果との相関により示される．

読解効率テストの比較対象テストとして，TOEIC 読解テストを利用する．TOEIC 読解テストは，テキストの理解度を問うテストであるが，その制限時間と読解対象テキスト量から，読解速度的能力もテスト結果に反映されていると考えられる．そこで，本稿は，TOEIC 読解テストを広義の読解効率テストとして位置づけた．また，TOEIC 自体の妥当性は TOEIC の開発実施機関である Educational Testing Service により確認されていることも比較対象とする理由である[5]．

3. 実験

テキストの読みやすさが読解効率テストの有効性にどのような影響を及ぼすかを確認するために，日本語母語話者を実験参加者として TOEIC（Test of English for International Communications）準拠テストのテキストを用いて実験を行った．

---

[1] 言語テストに対して，同一テストにより信頼性を推定すると学習効果により不適切であると考えられている[11]．読解テストとは異なるが，言語情報を排した文字列の読み上げにおいても，学習効果により約3%程度速度が上昇することが報告されている[12]．

[2] 目標関連妥当性以外に，基準関連妥当性や構成概念妥当性がテストの妥当性を示すと考えられている．基準関連妥当性は集団準拠テストに対する妥当性であり，テスト結果の正規分布によって示される．構成概念妥当性はテスト結果の弁別性によって示される．本実験において分析対象となる全てのテストが基準関連妥当性と構成概念妥当性（3 クラスの弁別性）のどちらも有していることを確認できた．

実験参加者は，事前に TOEIC を受験していることを条件とし，参加報酬を支払うことを明示して募集した日本語母語話者 107 名であった．全参加者のうち，実験で得られたデータに不備が発見された 5 名は分析対象外とした．最終的に，102 名が有効参加者となった．

有効参加者 102 名の TOEIC の読解テストのスコアの平均点は 311.1 点，標準偏差は 99.8 であった．また，最低点は 105 点，最高点は 470 点であった[3]．TOEIC スコアは，聴解テストのスコアと読解テストのスコアに分けられるが，本稿では読解テストの妥当性を検証することが目的であるため，読解テストのスコア（以下，TOEIC 読解スコアと呼ぶ）との比較を行った．

実験における読解対象テキストとして，TOEIC 準拠の問題集から抜粋した 42 テキストを利用した．実験用の 42 テキストを 6 通りのテキストセット（T1～T6）に分けた．各テキストセットに参加者をランダムに振り分けた．各テキストセットの平均参加者数は 17 名，標準偏差は 1.3 であった．最小参加者数は 16 名，最大参加者数は 19 名であった．

テキストの読みやすさを決める指標に関しては多くの議論が提出されている[6]．本稿では，読みやすさを示す指標として様々な場面で利用されているいわゆるリーダビリティスコアを用いる．リーダビリティスコアはその算出が容易であることから，授業において実際に運用することが可能であることが利点である．このリーダビリティスコアにも様々なスコア算出式が提案されているが，今回は単語長と文長といった非常に単純な指標から式（2）により算される Flesch Reading Ease[7] を用いることにした[4]．

スコア=206.835-(1.015×ASL)-(84.6×ASW)　（2）

ASL: 平均文長（一文の平均単語数）

ASW: 平均単語長（一単語の平均音節数）

実験用テキストの読みやすさであるリーダビリティスコアを算出した（表 1）．テキストセット全体のリーダビリティスコアの平均値は 58.29，標準偏差は 4.00 であった．また，最低スコアは 52.36，最高スコアは 62.30 であった[5]．

[3] 有効参加者の TOEIC 受験時期は，2002 年から 2005 年と 3 年にわたる．受験時期を年毎にみると，2002 年が 12 名，2003 年が 24 名，2004 年が 56 名，2005 年が 10 名であった．実験は 2005 年 3 月に行った．

[4] 本稿は読みやすさの指標として Flesch Reading Ease を採用したが，その妥当性は検討の余地がある．例えば，英語学習者用の算出式や語彙情報（頻度や親密度など）も考慮するべきかもしれない．

[5] Flesch Reading Ease は，読みやすさを 0 から 100 までのスコアで表す．スコア 60 から 100 までがおおよそ小学生・中学生程度のテキスト，スコア 30 から 60 までが高校生・大学生程度のテキスト，0 から 30 までが大学院生程度のテキストと考えられている．

表 1：リーダビリティスコア：T1～T6

テキストセット	N	リーダビリティスコア	
		平均値	標準偏差
T1	7	62.30	25.99
T2	7	60.68	12.51
T3	7	52.36	17.26
T4	7	57.77	13.08
T5	7	55.02	13.50
T6	7	61.61	14.38

各テキストセットを一つのテストとしてみなす．この場合，それぞれのテストにおいて参加者には同一のテキストが割り当てられることから，それぞれのテキストセットにおける有効性が同一テストにおける読解効率テストの有効性と考えられる．したがって，同一テストの有効性は，T1～T6 の各テキストセットにおいて検証する．

これら 6 つのテキストセットをリーダビリティスコアによる分散分析を行った結果，T1, T2, T4, T6 のテキストセットが同程度のリーダビリティスコアを有するテキストセットであることがわかった（$F=1.64$, $p=0.18$）．そこで，これら 4 つのテキストセットを一つのテスト（同程度テスト）とする．そして，テキストセット間のリーダビリティスコアに有意な差が確認できた T1 と T3 を異なりテストとする．これらのテストにおいて読解効率テストの有効性を検証する．

表 2：各テストにおけるテキストセット

同程度テスト	異なりテスト
T1	**T1**
T2	**T3**
T4	**-**
T6	**-**

参加者のテキスト読解時間とテキストの内容に関する設問に対する解答を，コンピュータ上の読解時間・解答記録ツール[8]を用いて記録した．

読解効率テストの有効性を三種類のテスト（同一・同程度・異なりテスト）において検証する．テストの有効性は，統計的検定によりテストの有効性を検証する古典的テスト理論[4]に基づいて検証した．まず，同一テストにおける読解効率テストの有効性を検証する．次に，同程度テストや異なりテストにおける読解効率テストの有効性を検証する．その結果に対して，同程度テストや異なりテストの有効性を比較する．もし，同程度テストの有効性が同一テストと同等であることが確認できれば，異なるテキストであってもリーダビリティスコ

アによって統制することにより同一テキストとみなすことができると考えられる.

4. 実験結果と考察

4.1. テストの有効性：同一テスト

同一テストにおける信頼性をクロンバック α 信頼性係数により検証を行う．同一テスト T1～T6 におけるクロンバック α 信頼性係数の算定の結果，信頼性係数は表 3 に示すように 0.80～0.92 であった.

表 3：同一テストのクロンバック α 信頼性係数

	N	読解効率（eWPM）		α 係数
		平均値	標準偏差	
T1	17	50.80	30.50	0.80
T2	19	61.70	34.87	0.87
T3	16	58.23	36.67	0.83
T4	18	54.85	39.36	0.91
T5	16	67.48	40.48	0.92
T6	16	48.33	35.30	0.80

古典的テスト理論では，信頼性係数が 0.7 以上あれば信頼できるテストであると考えられている. 読解効率データでは，全ての同一テストにおける平均信頼性係数が基準値の 0.7 を上回っている. したがって，読解効率テストの信頼性が同一テストにおいて確認できたといえる.

次に，同一テストにおける妥当性を読解効率データと参加者の TOEIC 読解スコアとの相関係数により検証を行う．スピアマンの順位相関係数の有意性検定を同一テスト T1～T6 において行った結果，相関係数は表 4 に示すように 0.68～0.87 であり，比較的強い相関関係が確認できた．また，T4 と T5 においては統計的に有意な強い相関関係が確認できた.

先行研究[9]，[10]によると，理解度と読解速度の間には相関係数が 1.00 に近い値ではないが，正の相関が見られるという. 実験条件は異なるがこれらの研究を参考に, TOEIC 読解スコアと読解効率データの間に統計的に有意な正の相関係数があれば，読解効率データが妥当性を有するとする. 実験の結果得られた読解効率データでは，T4 と T5 において統計的に有意な正の相関関係が確認できた.

表 4：読解効率と TOEIC 読解スコアの相関係数

	N	相関係数
T1	17	0.68(0.0025)
T2	19	0.69(0.0012)
T3	16	0.63(0.0085)
T4	18	0.87**
T5	16	0.81**
T6	16	0.68(0.0038)

**p<0.0001

T4 と T5 を除くテキストセットにおいては比較的強い相関関係が確認できるが，統計的に有意ではなかった．本実験ではデータ数が 20 以下と小規模であるため，はずれ値の影響なども大きいと考えられる. 読解効率テストの妥当性を検証する際に必要なはずれ値の処理や事例数は今後の課題とする.

これらの結果から，本稿の行った読解効率テストにおいて信頼性と妥当性による観点から同一テストの有効性が T4 と T5 において確認できた.

4.2. テストの有効性：同程度・異なりテスト

同程度テストと異なりテストにおける信頼性をクロンバック α 信頼性係数により検証を行う. それぞれのテストにおけるクロンバック α 信頼性係数の算定の結果，信頼性係数は表 5 に示すように，同程度テストが 0.81，異なりテストが 0.78 であった.

表 5：同程度・異なりテストのクロンバック α 信頼性係数

	N	読解効率（eWPM）		α 係数
		平均値	標準偏差	
同程度テスト	70	54.23	34.46	0.81
異なりテスト	33	54.4	33.77	0.78

どちらのテストも信頼性係数の基準値である 0.7 以上である．したがって，読解効率テストの信頼性は，同程度テスト，異なりテストのいずれにおいても確認できたといえる.

信頼性係数の特徴として，一般に項目数や事例数が多くなると信頼性が高くなると考えられている. 本実験において，同程度テストと異なりテストの項目数は 7 テキストで同じであるが，事例数が異なる. そこで同程度テストにおいて，異なりテストの場合と事例数を同程度にするために，二組のテキストセットにおける信頼性を確認した（表 6）.

表 6：同程度テスト（テキストセット二組）のクロンバック α 信頼性係数

	N	読解効率（eWPM）		α 係数
		平均値	標準偏差	
T1-T2	36	56.55	33.26	0.84
T1-T4	36	52.88	35.32	0.82
T1-T6	33	49.60	32.87	0.72
T2-T4	37	58.36	37.20	0.85
T2-T6	35	55.56	35.62	0.83
T4-T6	34	51.78	37.56	0.84

実験の結果，T1-T6 の組み合わせでは，同程度テストと異なりテストの場合の信頼性係数よりも

低くなったが，全ての組み合わせにおいて有効な信頼性係数が確認できた．このことから同程度テストにおけるテキストセットは，二組であっても信頼性が確認できたといえる．

次に，同程度テストと異なりテストにおける妥当性をそれぞれのテストにおける読解効率データと参加者の TOEIC 読解スコアとの相関係数により検証を行う．同程度テストと異なりテストにおける相関係数を算定した結果，相関係数は表 7 に示すようにそれぞれ 0.71 と 0.69 であり，統計的に有意な比較的強い相関関係が確認できた．

表 7: 同程度・異なりテストの読解効率と TOEIC 読解スコアの相関係数

	N	相関係数
同程度テスト	52	0.71**
異なりテスト	33	0.69**

**p<0.0001

どちらのテストも統計的に有意な正の相関係数を示している．したがって，読解効率テストの妥当性は，同程度テスト，異なりテストのいずれにおいても確認できたといえる．

信頼性係数と同様に相関係数も事例数による影響が大きいと考えられるため，同程度テストにおいて，二組のテキストセットにおける相関性を確認した（表 8）．実験の結果，全ての組み合わせにおいて統計的に有意な正の相関係数が確認できた．このことから同程度テストにおけるテキストセットは，二組であっても妥当性が確認できたといえる．

表 8：同程度テスト（テキストセット二組）の読解効率と TOEIC 読解スコアの相関係数

	N	相関係数
T1-T2	36	0.68**
T1-T4	36	0.79**
T1-T6	33	0.68**
T2-T4	37	0.76**
T2-T6	35	0.61**
T4-T6	34	0.75**

**p<0.0001

実験の結果，同程度テストと異なりテストのいずれにおいても統計的に有意な正の相関関係が確認できた．したがって，本稿の行った読解効率テストにおいて同程度テスト・異なりテストのどちらも信頼性と妥当性による有効性が確認できたといえる．

5. まとめ

本稿は，読解効率テストの有効性がテストに含まれるテキストにより異なるのかどうかを検証した．読解効率テストの有効性を，(1) 同一テキストにより構成されるテスト（同一テスト），(2) 同程度の読みやすさの異なるテキストにより構成されるテスト（同程度テスト），(3) 読みやすさもテキストも異なるテキストから成るテスト（異なりテスト）において分析を行った．分析の結果，テストの信頼性はテストの種類に関係なく有効であることが確認できた．一方，妥当性は同一テストの場合に統計的に有意でない相関関係が確認された．同一テストにおける事例数は 16～19 に対して，同程度テストや異なりテストの事例数は 33～52 であった．このことから，読解効率テストに影響を及ぼす要因として，読みやすさよりもテストに含まれる事例数の影響が大きいと考えられる．

今後，本研究の手法により収集した読解効率データの詳細な分析を行う．信頼性係数の分析に関しては，項目数や参加者数が信頼性係数にどのような影響を及ぼすかを調査する．目標関連妥当性の分析では，TOEIC 読解能力によるクラス分けを行い，各クラスの相関値を比較する．また，読みやすさの指標もリーダビリティスコア以外の指標を用いてその効果を調査する．

参考文献

[1] M. D. Jackson and J. L. McClelland, Processing determinants of reading speed, Journal of Verbal Learning and Verbal Behavior, vol. 108, pp. 151-181, 1979.

[2] R. Day, and J. Bamford, Extensive Reading in the Second Language Classroom, Cambridge University Press, Cambridge, 1998.

[3] 高梨庸雄, 卯城祐司 英語リーディング事典, 研究社, 東京, 2000.

[4] J. D. Brown, Testing in Language Programs, Prentice-Hall, Upper Saddle River, NJ., 1996.

[5] The Chauncery Group International, Ltd., TOEIC Technical Manual, The Chauncery Group International, Ltd., Princeton, NJ., 1998.

[6] A. H. Urquhart and C. J. Weir, Reading in a Second Language: Process, Product and Practice, Longman, London, 1998.

[7] R. Flesch1, The art of Readable Writing. Hamper, New York, 1949.

[8] 吉見毅彦, 小谷克則, 九津見毅, 佐田いち子, 井佐原均, “英語学習者の読解能力推定のための読解時間測定法, ”教育システム情報学会誌 vol.22, no.1, pp.24-29 2005.

[9] K. Kitao, Japanese colleage students’ English ability, MS., 1995.

[10] 門田修平, 野呂忠志編著, 英語リーディングの認知メカニズム, くろしお出版, 東京, 2001.

[11] J. C. Alderson and S. WIndeatt, Computers and innovation in language testing, in Language Testing in the 1990s: The Communicative Legacy, pp. 226-235, Macmillan, London, 1991.

[12] A. J. Wilkins, R. J. Jeans, P. D. Pumfrey, et al., Rate of reading test: its reliability, and its validity in the assessment of the effects of coloured overlays, Ophthalmic and Physiological Optics vol. 16, pp.491-497, 1996.