# 関西大学英語入試問題データの分析 : テスト理論の活用を目指して

| | |
|---|---|
| その他のタイトル | An Analysis of English Language Entrance Examination Data from Kansai University : Applying Language Testing Theory |
| 著者 | 水本 篤, 脇田 貴文, 名部井 敏代 |
| 雑誌名 | データ分析の理論と応用 = Bulletin of data analysis of Japanese Classification Society |
| 巻 | 6 |
| 号 | 1 |
| ページ | 21-29 |
| 発行年 | 2017 |
| URL | http://hdl.handle.net/10112/13021 |

データ分析の理論と応用　Vol. 6, No. 1 (2017), 21–29

論　　文

# 関西大学英語入試問題データの分析

## ——テスト理論の活用を目指して——

関西大学　水　本　　篤
関西大学　脇　田　貴　文
関西大学　名部井　敏　代

**要　旨**　大学入試は受検者の今後に大きな影響を与えるハイステークス・テストであるが，一般的に情報の開示が難しいため，その品質保証のためにテスト理論を活用してテスト自体の改善を行うという取り組みはこれまでに報告されていない．そこで本研究では，現行の関西大学英語入試問題実施データを分析し，その構造的妥当性を検証した．また，受検者の各問題セクションの合計得点が，テストの総合計点にどのように寄与しているかを探った．多次元項目反応理論による分析の結果，一般因子である英語総合能力とセクションや英文素材を反映した下位領域特有の因子による双因子モデルが支持され，現行の英語入試問題形式の構造的な側面の妥当性が確認された．また，特定セクションの合計得点が英語入試における受検者の能力弁別に役立っていることが示された．同時に，現行の英語入試問題作成における課題も明らかになったため，今後の大学入試問題作成や分析でテスト理論をより活用していくことにより，テストの品質改善が可能であることが示唆された．

**キーワード：**入試，英語，構造的妥当性，多次元項目反応理論，テスト理論

## 1. はじめに

英語 4 技能(リーディング，ライティング，リスニング，スピーキング)を統合的に測定することの重要性から，大学入試で英語 4 技能を測定することができるテストの導入を推奨する動きが国内で広がっている (中央教育審議会, 2014)．各大学で独自に 4 技能試験を作成するということは，問題作成業務や運用上のコスト，そして，大規模テストでライティングやスピーキングのパフォーマンスを厳密に評価することは現実的には難しい．そのため，入試で 4 技能試験を導入する際には，外部で作成された英語能力試験を利用することになることも必要だろう．ただし，そのような英語能力試験の導入が，数年以内に全面的に行われるとは考え難いため，各大学の入試問題作成担当部門は，現行の入試問題を測定論的な観点から検討・改善し，学生の合否判定のためにより正確な情報を得る努力を継続していかなければならない．

そのような必要性にもかかわらず，日本の大学入試では，作問段階で改善のために試行テストを行ったり，過去に実施した問題を分析し，良問をアイテム・バンクとして活用したり，テスト

を等化するというような，言語テスト分野で推奨されている一連の流れはほとんど行われておらず，その背景には「日本のテスト文化」があるとしばしば指摘される (柴山, 2008).

大学入試は受検者の今後に大きな影響を与えるハイステークス・テストであり，入試実施大学が開示できない情報も多く含まれており (倉元・西郡・木村・森田・鴨池, 2009)，項目レベルでの分析が行われない実情も理解できる．しかし，テストの品質管理と，より公平かつ正確な選抜を行えるテストの作成，実施，分析の必要性を考えると，測定論的観点を踏まえたテスト理論の利用が非常に重要である (宇佐美, 2016).

そこで本研究では，大学入試問題作成や分析でテスト理論をより活用していくために，現行の関西大学英語入試問題実施データを分析し，その構造的妥当性を検証した．また，受検者の問題セクションごとの解答パターンが，能力の弁別にどのように寄与しているかを探った．そして，これらの分析により現行の英語入試問題形式の課題を明らかにし，今後の入試問題作成改善のヒントを得ることを目的とした.

## 2. 本 研 究

### 2.1. 対象とした英語入試問題の概要

関西大学の入学試験は毎年 10 日間(2 月に 8 日間，3 月に 2 日間)実施され，英語は日数分の問題セット(10 セット)が作成される．本研究で分析の対象とした英語入試問題は，2010 年以降に実施された入学試験のうちから，受検者 10,982 名分のローデータを含んだ 1 セット(本研究ではセット P と呼ぶ)を対象とした.

対象としたセット P の構成，形式，問題数，配点は表 1 に示す通りである．現行の関西大学英語入試問題では，測定しているスキルにはリスニング，ライティング，スピーキングは含まれておらず，リーディングのみの 1 技能を対象としている．セクションは 6 つに分かれており，1A は会話文の空所を補充する形式で，読解力と日常口語表現の知識を問うている．1B は英文の順序を復元する整序問題で，パラグラフ・談話構造の理解を問うている．2A と 2B は 800 語～1000 語の物語風のパッセージを用い，2A では意味理解にかかわる語や句の空欄補充，2B では概要理解の読解問題が出題される．そして，3A と 3B では，800 語～850 語の記述的・解説的なパッセージを用い，3A では言い換えなどによって詳細の理解力を測定し，3B では要点の理解力を問う読解問題になっている．英文素材(パッセージ)は，1A の会話文，1B の段落整序(付録 1 参照)，2A と 2B の物語風パッセージ，そして 3A と 3B の記述的・解説的パッセージの 4 つの素材が用いられている．問題はすべて多枝選択式で合計 50 問，解答時間は 90 分のテストである.

現行の関西大学英語入試問題において，典型的な読解問題であるセクション 2B，3A，3B では，選択枝数は 3 つになっている．4 択の多枝選択式が多い大学入試において，3 択の問題形式が採用されているのは，同じ問題形式の実質選択枝数を比較した，Shizuka, Takeuchi, Yashima, & Yoshizawa (2006) の研究結果を受けたもので，「まったく選択されることがない 4 つ目の選択枝を苦労して作成するよりも，質の良い 3 択形式にすることによって，4 択形式と比較して，難易度，弁別力の変わらない問題の作成に注力すべきである」という考えを具現化したものになっている.

**表 1. 分析の対象とした英語入試問題の構成，形式，問題数，配点**

| 大問(セクション) | | 内容 | 問題形式 | 問題数 | 配点 |
|---|---|---|---|---|---|
| 1 | A | 会話文問題 | 4 択 | 5 | 各 4 点<br>(計 200 点) |
| | B | 段落整序問題 | 文の順序を復元 | 6 | |
| 2 | A | 本文空欄補充問題(読解) | 4 択 | 15 | |
| | B | 概要理解問題(読解) | 3 択 | 7 | |
| 3 | A | 詳細理解問題(読解) | 3 択 | 10 | |
| | B | 要点理解問題(読解) | 3 択 | 7 | |

## 2.2. 分析方法

本研究における分析はすべて，R version 3.2.3 (R Core Team, 2015) を用いた．まず，現行英語入試問題の構造的妥当性について検討するために，本研究で対象としたセット P(平均正答率 71%)に含まれる 50 項目の因子構造の確認を，多次元項目反応理論(Multidimensional Item Response Theory; 多次元 IRT)を用いて分析した．従来の項目反応モデルでは一次元性(一因子構造)が仮定されているのに対し，多次元項目反応モデルは多次元(多因子)に拡張されたものであり，因子分析の文脈においては多因子モデルに対応する (小杉, 2014)．多次元 IRT を構造的妥当性の検証に用いた坂本 (2016) では，双因子モデル(bifactor model)を用いることによって，テスト項目全体に影響を与えている一般因子(general factor)と下位領域特有の構成概念に影響を与えているグループ因子(group factor)の両方をモデリングし，一次元性を仮定している通常の IRT では同定することができなかった多次元の構成概念をモデリングできることを明らかにした．

また，多次元 IRT を本研究で用いた大きな理由としてテストレットへの対応が挙げられる．テストレットとは，大問形式で構成されている項目群のことを指す (森・大森・繁桝，2011)．2.1 や表 1 からも分かる通り，現行英語入試問題は 4 つの英文素材(1A，1B，2A & 2B，3A & 3B)に対して，6 つのセクションで構成されており，項目の 1 つ 1 つが独立しておらず，従来の IRT で重視される，「1 つの項目の解答が別の項目の解答に影響を与えない」という，局所独立性(local independence)の仮定が厳密には満たされていない．本研究で多次元 IRT 分析に用いた，R パッケージの mirt (Chalmers, 2012) では，そのようなテストレットをモデリングに組み込むことができることから，多次元 IRT により，因子構造のより正確な推定が可能であると考えられた．

多次元 IRT に先立って，セット P の 50 項目に対して正解・不正解の 2 値データを使って項目合計相関を求めたところ，0.2 以下の項目がセクション 2A において 2 項目存在したため(正答率 18.8%と 34.2%)，この 2 項目を削除し，セクション 1B の段落整序問題の 6 問を 0〜6 点の範囲とする多値型の 1 項目とし，合計 43 項目を使って多次元 IRT を行った．

因子構造を MAP(Minimum Average Partial，最小偏相関平均)基準で調べたところ，因子数は 1〜4 となり，シンプルな因子構造であることがわかった．そのため，本研究では坂本 (2016) に倣い，(a) モデル A：すべての項目を使った 1 因子モデル，(b) モデル B：6 つのセクションの相関モデル，(c) モデル C：一般因子である「英語総合能力」と 6 つのセクションの双因子モデル，(d) モデル D：4 つの英文素材の相関モデル，(e) モデル E：一般因子である「英語総合能力」と 4 つの英文素材の双因子モデルの 5 つのモデル設定し，情報量規準と適合度指標の比較を行った．これにより，どのモデルがテストの構造を一番うまく説明できているかを検討した．

次に，多次元 IRT で比較したモデルに関連して，受検者の各問題セクションの合計得点が，テ

ストの総合計点にどのように寄与しているかを探るために，偏決定係数と共分散比を用いて比較を行った．また，ある特定のセクションの合計得点が同じ受検者でも，他のセクションの合計得点が違うパターンを示すことも十分考えられるため，受検者の各問題セクションの合計得点とテストの総合計点の関係を明らかにすべく，クラスター分析(ウォード法・ユークリッド距離)を行った．

## 3. 結果と考察

多次元 IRT によるモデル比較の結果を表 2 に示す．情報量規準は，AIC(Akaike's Information Criterion)，AICc(corrected AIC)，BIC(Bayesian Information Criterion)，SABIC(sample size adjusted AIC)，DIC(Deviance Information Criterion)の 5 つを用いた．適合度指標は，CFI(Comparative Fit Index)，TLI(Tucker-Levis Index)，RMSEA(Root Mean Square Error of Approximation)，SRMSR(Standardized Root Mean Square Residual)の 4 つを用いた．5 つの情報量規準と RMSEA と SRMSR は値が小さいほど，そして，CFI と TLI は値が大きいほど当てはまりがよいと考えられるため，情報量規準ではモデル C，適合度指標ではモデル E を支持する結果となった．

**表 2. 多次元 IRT によるモデル比較の結果**

| 指標 | モデル A | モデル B | モデル C | モデル D | モデル E |
|---|---|---|---|---|---|
| AIC | 481039.234 | 480426.451 | 479673.586 | 480503.379 | 479814.204 |
| AICc | 481040.772 | 480428.536 | 479677.713 | 480505.125 | 479817.845 |
| BIC | 481703.899 | 481200.676 | 480761.884 | 481211.868 | 480836.765 |
| SABIC | 481414.713 | 480863.822 | 480288.381 | 480903.614 | 480391.863 |
| DIC | 481039.234 | 480426.451 | 479673.586 | 480503.379 | 479814.204 |
| CFI | 0.974 | 0.983 | 0.981 | 0.982 | 0.987 |
| TLI | 0.973 | 0.982 | 0.979 | 0.981 | 0.986 |
| RMSEA | 0.018 | 0.015 | 0.016 | 0.015 | 0.013 |
| SRMSR | 0.020 | 0.024 | 0.041 | 0.018 | 0.019 |

モデル C は，一般因子である「英語総合能力」と 6 つのセクションの双因子モデルであり，モデル E は，一般因子である「英語総合能力」と 4 つの英文素材の双因子モデルであるため，一般因子がテストの総合計点に影響していることは自明であり，さらに，6 つのセクションの問題形式や，4 つの英文素材の影響を受けているということがわかり，現行の関西大学英語入試問題は「全体的な能力に加え，セクションや英文素材が反映された下位領域を測定しているテスト」であると言える．

このように，セクションや英文素材の影響が多次元 IRT により明らかになったため，各問題セクションの合計得点が，テストの総合計点に与える影響を調べた結果を表 3 に示す(表 1 の問題数や配点からもわかるように，このテストでは設問ごとに配点は違わないため，以下では 1 問 1 点で分析を行っている)．この結果から，セクション 2A がテストの総合計点にもっとも影響を及ぼしていることがわかる．セクション 2A は，他のセクションと比べて問題数も多いため(表 1 参照)，その影響は考慮すべきではあるが，問題形式としては，本文空欄補充問題(読解)であり，本文の内容を理解した上で，文脈にふさわしい語や句を補充する必要があるため，その解答には，よ

り高次な読解能力が求められる (Mizumoto, Ikeda, & Takeuchi, 2016).

**表 3.** 各問題セクションの合計得点とテストの総合計点の関係

| セクション | 平均値 | 標準偏差 | 相関係数 | 偏決定係数 | 共分散比 |
|---|---|---|---|---|---|
| 1A | 4.237 | 0.925 | 0.537 | 0.014 | 7.120 |
| 1B | 5.192 | 1.527 | 0.565 | 0.040 | 12.369 |
| 2A | 8.545 | 2.732 | 0.815 | 0.098 | 31.911 |
| 2B | 4.889 | 1.541 | 0.709 | 0.033 | 15.669 |
| 3A | 6.965 | 1.738 | 0.733 | 0.040 | 18.276 |
| 3B | 5.875 | 1.412 | 0.724 | 0.026 | 14.654 |
| 総合計点 | 35.703 | 6.974 | — | — | — |

たとえば，セクション 2A の合計得点が同じ受検者でも，他のセクションの合計得点が違うパターンを示すこともあるため，クラスター分析を行い，その結果を基に，セクション合計得点とテスト総合計点のパターンを図示したものが図 1 である．クラスター数はデンドログラム(付録 2)を見て，5 つが適切であると判断した．以下の図からわかるように，クラスター 1($n = 3,147$)の受検者はすべてのセクションで合計得点が高く，テスト全体の総合計点も高い($M = 43.07$)．対照的に，クラスター 5($n = 792$)の受検者はすべてのセクションで合計得点が低く，テスト全体の総合計点も低い($M = 21.68$)．入試形態や学部，受験科目によって基準は違うものの，クラスター 1 の受検者はおそらく合格となり，クラスター 5 の受検者はおそらく不合格と判定されることが予想される．クラスター 2($n = 3,640$)，クラスター 3($n = 1,741$)，そしてクラスター 4 ($n = 1,662$)の受検者では，セクション 1B の合計得点が高いクラスター 2 の受検者のテスト総合計点($M = 36.42$)が，同じセクション 1B で合計得点の低いクラスター 3 の受検者のテスト総合計点($M = 32.28$)よりも高くなっており，セクション 1B の合計得点が(段落整序問題であるため，採点方法の影響があるものの)英語入試における受検者の弁別に役立っていることがわかる．また，クラスター 4 の受検者はセクション 1B の合計得点が，テスト総合計点が高いクラスター 2 の受検者と同じであるが，クラスター 4 の受検者はセクション 2A，2B，3A，3B の得点が低いため，テスト総合計点も低くなっており($M = 30.43$)，これらのセクションの影響もあることがわかる．

## 4. 研究の成果と課題・今後の展望

本研究では，大学入試問題作成や分析でテスト理論をより活用していくために，関西大学の現行英語入試問題データを分析した．多次元 IRT によって，一般因子である「英語総合能力」と，セクションや英文素材を反映した下位領域特有の因子による双因子モデルが支持された．また，受検者の各問題セクションの合計得点が，テストの総合計点に与える影響を精査し，特定セクションの合計得点が英語入試における受検者の弁別に役立っている可能性が示唆された．

本研究を通して明らかになった課題としては，項目合計相関が低い項目があり，多次元 IRT ではこれらの項目を削除して分析したことからも，受検者にとって易しすぎる問題，難しすぎる問題も存在しており，これらの失敗・成功例を蓄積し，問題作成に十分反映できていないことが挙げられる．特に，能力の弁別に役立っているテスト項目は，再利用できないとしても，アイテム・

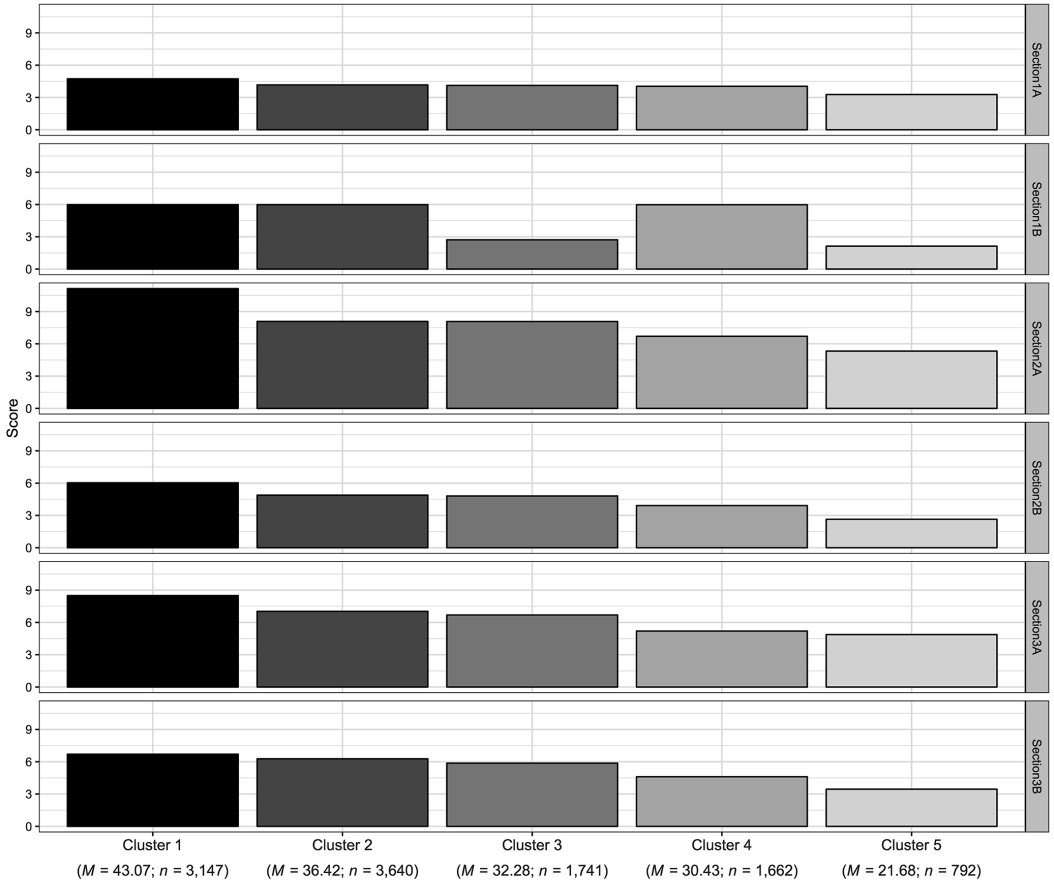



**図 1.** クラスター分析から得られたセクション合計得点とテスト総合計点のパターン

バンクの考え方を取り入れ，良い項目例と具体的な作成方法を蓄積していくことが可能である．そのためにはテスト理論の活用が欠かせない．入試の目的は合否判定であるが，本研究で示したように，テスト理論を活用することによって，合格・不合格だけの情報ではなく，テスト作成者が参考になる情報を得ることができる．そしてその情報を基に，テスト設計(スペック)の改善に活用していくことが可能になる．

テストの作問が職人たちによる芸術(アート)であるとすれば，そこにテスト理論による知見や分析結果を取り入れていくのは科学(サイエンス)である．より公平なテスト作成，実施，分析のためには，「アートとサイエンスの融合」をテストにかかわるものは常に目指していかねばならず，テスト作成者と分析者がともに協力して，より品質の高いテストを作成できる体制作りが必要である．

また，関西大学の現行英語入試問題は 1 技能(リーディング)しか測定しておらず，第二言語習得やテストの波及効果の観点からは好ましいとはいえない．パフォーマンス・テスト(ライティング，スピーキング)の測定の信頼性・妥当性や，実現可能性を考えると，学内で 2 技能以上のテストを作成することは難しいため，外部で作成された英語能力試験を一部取り入れていくような試

みも必要であり，そのようなテスト利用の妥当性についても検討していかなければならない．

入試のローデータを分析し，その後の問題作成業務に反映していくというような前例は国内ではあまり報告されていない．しかし，本研究の結果からもわかるように，継続的にテスト理論を利用した分析を行うことによって，より測定精度の高いテストの作成が可能になると考えられる．試行テストの実施やテストの等化，アイテム・バンクの考え方が前提となっていない大学入試においても，現行の問題形式による選抜は続いていく．そのため，このような取り組みは全国の大学において推奨されるべきである．

## 謝　辞

本研究を遂行するにあたり，関西大学入試センターから入試データの提供をいただいた．結果の公表については，関西大学心理学研究科倫理委員会，関西大学入試センター所長，関西大学長の承認を得た．また，査読者から本稿の改訂プロセスにおいて，的確かつ建設的なコメントをいただいた．ここに記して感謝したい．本研究の一部は JSPS 科学研究費補助金（科研費）16K13273，26370719，16H02051，26370719 の助成を受けて実施された．

## 参 考 文 献

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48. doi:10.18637/jss.v048.i06

中央教育審議会 (2014). 新しい時代にふさわしい高大接続の実現に向けた高等学校教育, 大学教育, 大学入学者選抜の一体的改革について. Retrieved from http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/__icsFiles/afieldfile/2015/01/14/1354191.pdf

小杉孝司 (2014). 項目反応理論. 小杉考司・清水裕士（編著）, M-plus と R による構造方程式モデリング入門 (pp.165–187). 北大路書房.

倉元直樹・西郡 大・木村拓也・森田康夫・鴨池 治 (2009). 選抜試験における得点調整の有効性と限界―合否入替りを用いた評価の試み―. 日本テスト学会誌, 4, 135–152.

Mizumoto, A., Ikeda. M., & Takeuchi, O. (2016). A comparison of cognitive processing during cloze and multiple-choice reading tests using brain activation. *ARELE (Annual Review of English Language Education in Japan)*, 27, 65–80.

森 一将・大森 拓哉・繁桝算男 (2011). 相関構造を仮定したテストレットモデルのベイズ推定―国立大学法人等の教育研究評価データへの適用―. 大学評価・学位研究, 12, 3–16.

R Core Team. (2015). R: A language and environment for statistical computing (Version 3.2.3) [Computer software]. Vienna, Austria. Retrieved from http://www.r-project.org/

坂本佑太朗 (2016). わが国の TIMSS2011 数学データにおける多次元 IRT を使った妥当性の検証について. 日本テスト学会誌, 12, 37–53.

柴山 直 (2008). 日本のテスト文化について. 人事試験研究, 208, 2–13.

Shizuka, T., Takeuchi, O., Yashima, T., & Yoshizawa, K. (2006). A comparison of three- and four-option English tests for university entrance selection purposes in Japan. *Language Testing*, 23, 35–57. doi:10.1191/0265532206lt319oa

宇佐美 慧 (2016). 測定・評価・研究法に関する研究の動向と展望―教育測定・心理統計の専門家の不足および心理統計教育の問題の再考と「専門家による専門家の育成」の必要性―. 教育心理学年報, 55, 83–100.

（2016 年 6 月 29 日受付　2016 年 10 月 3 日修正　2016 年 12 月 5 日修正　2016 年 12 月 5 日採択）

著者連絡先：〒 564–8680 大阪府吹田市山手町 3–3–35 関西大学外国語学部
水本 篤（Tel. 06–6368–0508）
E-mail: mizumoto@kansai-u.ac.jp

**付録1**

セクション 1B の指示文

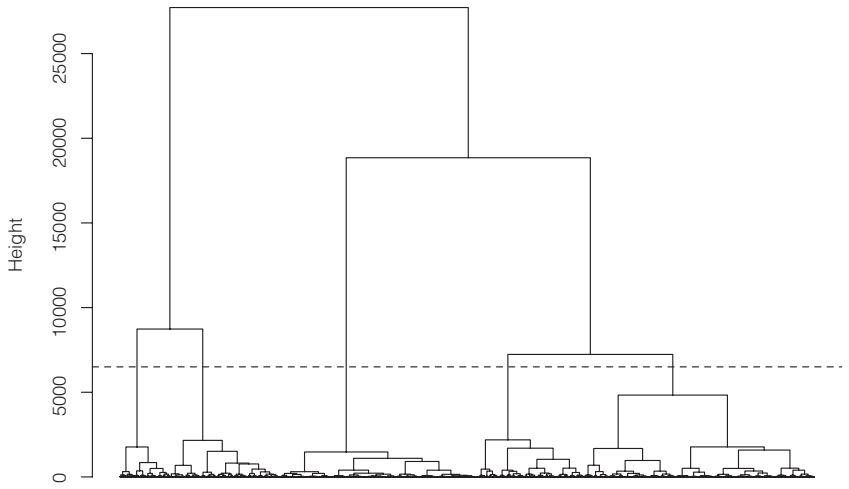
下の英文 A～F は，一つのまとまったパラグラフを，6 つの部分に分け，順番をばらばに入れ替えたものです．ただし，パラグラフの最初には A がきます．A に続けて B～F を正しく並べ替えなさい．その上で，次の(1)～(6)に当てはまるものの記号をマークしなさい．ただし, 当てはまるものがないもの(それがパラグラフの最後であるもの)については，Z をマークしなさい．

(1)A の次にくるもの
(2)B の次にくるもの
(3)C の次にくるもの
(4)D の次にくるもの
(5)E の次にくるもの
(6)F の次にくるもの

［以下，A～F の英文が続く］

**付録2**

クラスター分析で得られたデンドログラム



Bulletin of Data Analysis of Japanese Classification Society 6 (1) (2017), 21–29

# An Analysis of English Language Entrance Examination Data from Kansai University

—Applying Language Testing Theory—

**Atsushi Mizumoto*, Takafumi Wakita and Toshiyo Nabei**

Kansai University

**Abstract**

In this study, we analyzed an English language entrance test of Kansai University administered in the past to validate its structural aspect. We also investigated the extent to which each section of the test had an influence on the overall test score. By applying the multidimensional item response theory (MIRT), we found that a bifactor model with a general factor and lower-level factors reflecting the differences of sections and reading materials fit the data better than other models. In addition, it was confirmed that a certain section in the test effectively performed to distinguish the ability of test takers. These results suggest that the test retains desirable traits as a measurement instrument used for entrance examination purposes in terms of the structural aspect of validity and assessment of test takers' ability. At the same time, this study points to the importance of applying language testing theory in the process of creating, administering, and analyzing a high-stakes test. Implications of employing test theories are discussed in view of the current situation of entrance examination administration in Japan.

**Key words:** entrance examination, English, structural validity, multidimensional IRT, test theory

*Corresponding author

E-mail address: mizumoto@kansai-u.ac.jp (Atsushi Mizumoto)

Received June 29, 2016; Received October 3, 2016; Received December 5, 2016; Accepted December 5, 2016.