共通項目数が等化の精度に及ぼす影響

―大規模学力テストデータを用いた探索的研究―

泉　　毅*，山野井　真　児*，山　田　剛　史**，
金　森　保　智***，対　馬　英　樹***

（平成23年 6 月14日受付，平成23年12月 8 日受理）

Investigation of the equating accuracy under the influence of common item sizes：

Application of IRT test equating to the large-scale high school proficiency test data

IZUMI Tsuyoshi *, YAMANOI Shinji *, YAMADA Tsuyoshi **,
KANAMORI Yasutomo ***, TSUSHIMA Hideki ***

The purpose of this study is to examine the accuracy of IRT(Item Response theory) test equating. In this study, we examined the four features of test equating: (1) numbers of common items, (2) item discrimination parameter, (3) sample size, and (4) heterogeneous examinee groups.

We used the empirical test data that was provided from Benesse Corporation. We adopt Root Mean Square Error (RMSE) as the index of equating accuracy. Also, in cases where there is a gap in the group mean of ability between the examinee groups, equating accuracy get especially worse.

As a result, we concluded that at least 6 common items are required for the adequate accuracy of test equating.

Key Words : Item response theory, test equating, large-scale high school proficiency test

Ⅰ　問題と目的

Ⅰ-1　背景

我が国の学校現場で用いられるテストや，入学試験などのほとんどのテストは，古典的テスト理論(Classical Test Theory: CTT)に基づいている。古典的テスト理論は数学的なモデルとして比較的単純で，現実のテスト場面で使いやすいため，広く適用されている。(孫, 2002)[1]

大学入試センター試験も古典的テスト理論に基づくテストであるが，そのテストで得られた得点は，教科ごとにそのまま合計され，大学入試の選抜の場面に利用されている。

しかし，学力の経年変化を検討する場合，大学入試センター試験の毎年の平均点を用いることはできない。年度ごとに，平均点が変化したとしても，テスト得点は，テスト項目の難しさによる変化なのか，受験者の能力による変化なのか判断できない。これが，古典的テスト理論における主な問題点である。

この問題を解決できるのが，項目応答理論(Item Response Theory: IRT)である。項目応答理論では，テスト項目に関する情報(難易度など)と，受験者に関する情報(受験者の能力)とを分離して推定することができる。この性質により，異なるテストの得点が比較可能になるのである。例えば，全世界規模で実施される英語試験の 1 つであるTOEFL(Test of English as a Foreign Language: ETS)では，受験した時期によって問題が異なるにも関わらず，テストの得点は同じ意味をもって相互に比較できるようになっている。これは，TOEFLが項目応答理論に基づいて作られているからである。

Ⅰ-2　項目応答理論のモデル

項目応答理論でよく用いられるモデルである 2 パラメータ・ロジスティック・モデル(以下, 2PLMと記述する)について説明する。

このモデルは，以下の式で表される[2]。

$$P_j(\theta) = \frac{1}{1 + \exp\left(-Da_j(\theta - b_j)\right)}$$

θ は，項目応答理論によって求められた能力パラメー

*　岡山大学大学院教育学研究科(Master Program student of Graduate School of Education, Okayama University)
**　岡山大学(Okayama University)
***　ベネッセコーポレーション(Benesse Corporation)

タである。$P_j(\theta)$は，θの能力パラメータを持つ受験者が，テスト項目 j に対して応答した時に求められる正答確率を表す。b_jはテスト項目 j の項目困難度パラメータである。 Dは1.7の定数である。a_jはテスト項目 j の項目弁別力パラメータである。

このように，項目応答理論では，受験者の能力パラメータと，項目の特性を表す項目パラメータとを分けたモデルとなっている。

項目応答理論の他のモデルとして，多肢選択式問題において答えが分からずに当て推量で正答する確率をモデルに組み込んだ 3 パラメータ・ロジスティック・モデル(以下，3PLMと記述する)，項目困難度パラメータのみを扱う 1 パラメータ・ロジスティック・モデルがある。3PLMは当て推量が考えられる場合にモデルとデータがよく適合するという利点を持つ。しかし，大友(1996)[3] では，3PLMでは標本数が最小で1000から2000必要であること，当て推量が考えられることを前提としていることから，限定的な状況でのみ使えるモデルとなっている。今回の研究では，より一般的と考えられる 2 パラメータ・ロジスティック・モデルを用いて研究を行っている。

Ⅰ-3　等化の意味とその手法

資格や検定のためのテストは，毎回同じ測定領域を持つが，異なった問題，異なった受験者によって成り立っている。この場合，異なるテスト間の点数の解釈が問題となる。例えば，4 月に実施された英語のテストで100問中50問正解したとする。続いて，同じ領域の英語のテストの問題を 9 月に実施し，100問中60問正解した場合，本当に能力が上がっているのかを明らかにすることはできない。なぜなら，テスト間の困難度の差が考慮されていないので，9月の問題が易しかったために正答数が上がったと解釈できてしまうからである。

このような問題を解決するためには，それぞれのテストの得点を同一の尺度上の値に変換して表すことが必要である。これが，テストの等化である。

等化は，古典的テスト理論によって行うこともできる。しかし，古典的テスト理論においては，問題点が指摘されている。大友(1996)[3]では，素点が等化される場合には，等化するテストが平行であるか，テストの信頼性が全く等しいものでなければ，公平性条件の必要条件を満たすことはできないとしている。平行テストとは，同一の尺度に変換したあとで，平均値と標準偏差が等しく，しかもいかなる外部基準との相関も等しい 2 つのテストを指す。公平性条件とは，同一の能力を持っている受験者集団にとって，テスト X の得点分布と，等化されたあとのテスト Y に関する得点分布とは，同じものでなければならないということである。

この条件は，現実のテストの実施場面を考えると厳しいものとなっている。

これに対し，項目応答理論では，古典的テスト理論の等化の条件について考慮せず，等化を行うことができる。

項目応答理論によって等化する場合，テスト間に同一の受験者を設定するか，テスト間に同一の問題項目を設置する必要がある。前者を共通受験者デザインといい，後者を共通項目デザイン，または係留テストデザインという。ただし，共通受験者デザインでは，両方のテストを受ける受験者の学習効果や疲労の問題がある。このため，実際にテストの等化を行う場面では，共通項目デザインの方が多く用いられている。

項目応答理論における，代表的な等化の方法としては，4 つのものがある。同時尺度調整法，困難度固定法，困難度等化法，特性曲線等化法である。(Petersen ,Kolen,& Hoover, 1989)[4]今回は，同時尺度調整法を用いる。同時尺度調整法は，1 回の推定で全ての作業が終了するが，他の方法は項目母数の推定をテストごとに行った上で改めて等化をする必要があり手続き的に複雑になるためである。(藤森, 1997)[5]

同時尺度調整法は，2 つのテストに含まれる項目の各パラメータが一度に推定される。このとき，共通項目や共通受験者が 2 つのテストのデータをつなぐ役割を果たし，得られる結果は両テストに共通の尺度上のものになるというものである。

Ⅰ-4　等化に関する問題と先行研究

ここで，共通項目デザインによるテストを新たに作成しようと考えた場合，共通項目をテスト全体の中でどれだけ設ければよいのかという問題が生まれる。豊田(2002)[6]によると，共通項目数の目安として最低 5 つ必要であるとしている。しかし，これは絶対的な基準であるとはいえない。

等化の精度という観点からは，共通項目数は多いほうが良いとされる。しかし，共通項目が多すぎる場合，等化するテストがほとんど同じものになるために，複数のテストに分ける意味合いが薄れる。 2 つの異なるテストを 1 つの尺度にのせることができるという利点を得るには，共通項目数は，等化の精度を維持したうえで，少ない方がよいということになる。

藤森(1997) [5]，藤森(1998) [7]では，シミュレーションによって垂直的等化による能力パラメータの精度について検討した結果，共通項目数が多くなるにつれて等化の成績が改善すること，共通項目として，6 〜 8 個の共通項目が必要であることを報告した。しかし，このテストのデータは，項目パラメータと受験者の能力パラメータの真値が分かっているという想定でのシミュレーションによる分析となっている。そのため，実際に得られたテストデータにおいては，この目安が適用されない可能性が

考えられる。

一方，前川ら(2002)[8]，熊谷ら(2007)[9]のように実際のテストデータを用いて等化を行った研究もあるが，これらは等化の精度に着目した研究ではない。

以上の先行研究を踏まえ，本研究では実際のテストデータを用いて，共通項目数が等化の精度に及ぼす影響について分析を行う。

ただし，共通項目数以外にも等化の精度に影響を与える要因もある。藤森(1998)[7]では，全体の項目数，共通項目の項目弁別力，受験者数，受験者集団の能力値差がこれに関係すると述べられている。

本研究では共通項目数に加え，受験者数，共通項目の項目パラメータ，受験者集団の能力値差，の 4 つを等化の精度の要因として分析した。

Ⅱ　方法

Ⅱ-1　分析対象

本研究では，(株)ベネッセコーポレーションで実施された，高校一年生の基礎学力を測定するための多肢選択形式のテストを使用する。本研究で使用したテストデータは，2008年と2009年度に実施された，英語，国語の 2 教科のテストデータである。なお，2008年と2009年で同一のテスト項目が用いられている。

表2.1 用いたテストデータ

	2008 年度 英語	2009 年度 英語	2008 年度 国語	2009 年度 国語
項目数	56	56	43	43
受験者数	35779	41381	34773	40270
平均	35.020	34.063	26.843	26.839
標準偏差	9.249	9.288	6.714	6.856

Ⅱ-2　基本的な分析手順

本研究では，3 つの分析を行った。(1)共通項目の項目弁別力に着目した分析，(2)受験者数に着目した分析，(3)受験者集団の能力値差に着目した分析，である。ただし，各分析に共通して，共通項目数が減少した場合，どの程度等化の精度が低下するのかについて検討している。これらの 3 つの分析に共通する分析手順について説明する。

まず，受験者数を一定にする。2008年度と2009年度のテストデータから，それぞれ同一の受験者数分のデータを無作為に抽出した。

次に，非共通項目を作成する。非共通項目とは，共通項目ではない項目を指す。本研究で用いる 2 つのテストデータは，全てが共通項目となっている。これを簡易化して説明するために図2.1に示した。

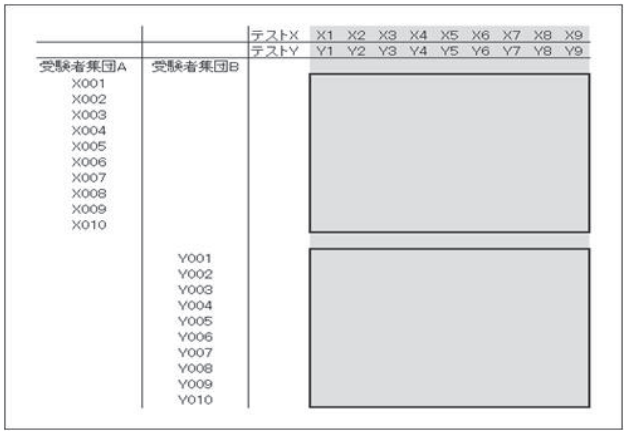

図2.1 全てが共通項目である元データ

この図は，10人の受験者集団 A が 9 項目のテストX を，10人の受験者集団 Bが 9 項目のテストYを受験したテストデータをあらわしている。テスト X とテスト Y は，全て同一の項目となっている。

共通項目デザインによる等化を行うテストデータとして，全てが共通項目であるというのは考えにくい。なぜなら，全ての項目が同一ならば，2 つのテストの受験者を同一の集団とみなして分析することができるため，等化の必要がないからである。そのため，テストデータを部分的に欠損させ，2 つのテスト間に非共通項目を設定した。共通項目がテスト全体の半分を占めるようにテストデータの削除を行った。これを図式したものとして，図2.2に示す。

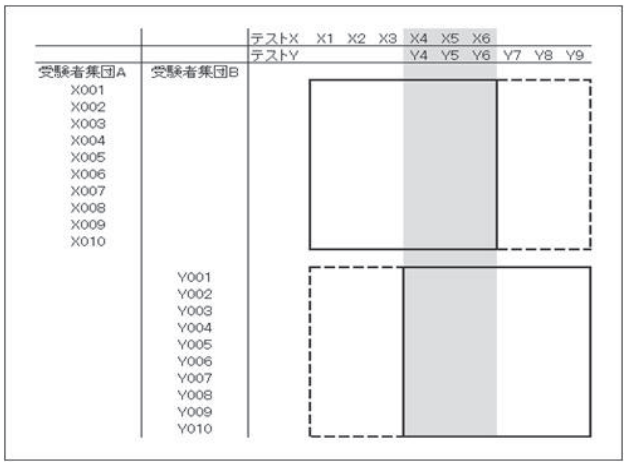

図2.2非共通項目を設けるように テストデータを削除した場合

図2.2では，点線で囲まれたテストデータを削除することで，非共通項目を作っている。この操作で得られたテストデータを，ここでは基準テストと呼ぶ。基準テストを等化し，受験者 i の能力パラメータを求める。ここで得られる受験者の能力パラメータを θ_i^*とする。なお，能力パラメータの分析には，BILOG-MG(Zimowski, Muraki, Mislevy, & Bock, 2003)を用いた。

次に，基準テストよりも共通項目数の少ないデータを作成する。その際，全体の項目数を変えずに，共通項目の数を減らしている。全体の項目数を変えないのは，項目数が変わることによる等化の精度への影響を統制するためである。

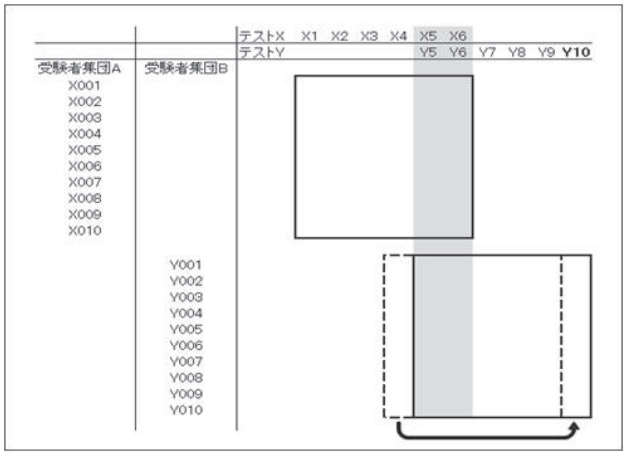

図2.3　共通項目の一部を非共通項目とみなした場合

図 2.3では，Y年度テスト項目のY4の項目を，新たに設けた項目Y10としている。このことにより，X4とY4は本来共通項目であるが，Y4をY10として，非共通項目とみなして分析することになる。これにより，全体の項目数を変えずに，共通項目の数を減らすことができる。以上の工程で共通項目数を減らし，共通項目数が2,4,6,8,10,12の場合のテストデータを作り，それぞれ等化を行う。また，等化後に得られるそれぞれの受験者iの能力パラメータを$\hat{\theta}_i$とする。

最後に，等化の精度を算出する。等化の精度が悪くなると，基準テストから得られた受験者 i の能力パラメータ θ_i^*と，共通項目数を減らした場合のテストデータから得られた受験者 i の能力パラメータ $\hat{\theta}_i$には差が生じると考えられる。θ_i^*と $\hat{\theta}_i$の差の指標として，De Ayala, Plake & Impara(2001)[10]を参考に次式とした。

$$\mathrm{RMSE}=\sqrt{\frac{\sum(\theta_i^*-\hat{\theta}_i)^2}{n}}$$

RMSEとは，平均二乗誤差(Root Mean Square Error)のことである。これは，比較したい値同士の間に平均的にどれだけ差があるかを示したものである。

式中の n はテストの総受験者数を指す。また，θ_i^*は，基準テストで得られた受験者iの能力パラメータを，$\hat{\theta}_i$は共通項目数を減らしたテストデータから得られた受験者 i の能力パラメータを表す。これを各受験者で差を取り，2乗したものを平均した値の平方根がRMSEとなる。

このRMSEが，共通項目の項目弁別力，受験者数，受験者集団の能力値差によって，どのように変化するのか 3 つの分析によって確認する。

以上の手順をまとめると，次のようになる。

手順 1：それぞれの年度の受験者数を固定する。

手順 2：非共通項目を設け，基準テストを設定する。

手順 3：基準テストよりも，共通項目数の少ないデータを設定する。

手順 4：それぞれのデータから得られた能力パラメータから等化の精度を求める。

Ⅱ-3　共通項目の項目弁別力に着目した分析の分析手順

この分析では，共通項目の項目弁別力の高低が等化の精度にどの程度影響を与えているのかについて検討する。ここでは，それぞれのテストデータの受験者数を2000人として分析を行った。これは，大友(1996)[3]にある3PLMに必要とされる最小標本数が1000から2000であることを根拠にしている。この研究では，2PLMによる分析を行うが，標本数2000であれば2PLMの分析として十分な標本数が得られると考えられる。

Ⅱ-2節における手順 3 では，共通項目を非共通項目にするよう，テストデータの整形を行っている。しかし，どの項目から非共通項目にするかによって，等化の精度が変わることが予想される。

そのため，あらかじめ元々のテストデータを分析し，テストの項目弁別力を得た上で，どの共通項目から非共通項目にするのかを検討した。

項目弁別力を得るために，1 つのテストデータあたり無作為に25000人ずつ抽出し，2PLMによる分析を行った。

得られた項目弁別力の小さい順に，共通項目を並び替える。このテストデータから，項目の弁別力の違いをもとに，3 種類のテストデータを作る。

まず，共通項目として，項目弁別力の低いものを残すため，項目弁別力の高いものから順に 2 つ共通項目を減らす項目弁別力の低グループを作る。

次に，共通項目として，項目弁別力の高いものを残すため，項目弁別力の低いものから順に 2 つ共通項目を減らす項目弁別力の高グループを作る。

最後に，共通項目として，項目弁別力の中程度のものを残すため，共通項目の項目弁別力の最も高いものと最も低いものを順に 1 つずつ共通項目減らす，項目弁別力の中グループを作った。

このそれぞれについて，共通項目数がテスト全体の半分を基準テストの共通項目数とし，共通項目数を12,10,8,6,4,2と変化させた場合，受験者集団の能力パラメータがどのように変化するのかについて検討する。基準テストの共通項目数は，英語は20項目，国語は15項目となっている。共通項目数がテスト全体の半分を占める場合を基準としたのは，現実のテスト実施場面で考えられる共通項目数として十分大きい数であると考えられる

ためである。ここでは，RMSEについて計18のケースを検討することになる。

Ⅱ-4　受験者数に着目した分析の分析手順

この分析では，できるだけ項目弁別力の影響を受けないように統制を加え，受験者数が等化の精度に与える影響について考える。このためには，Ⅱ- 3 節で紹介した，共通項目に残す項目の項目弁別力の高，中，低のいずれかにそろえる必要がある。ここでは，中程度の方法に揃えることにした。これは，項目弁別力が等化の精度に特に大きな影響を与える場合，項目弁別力の高，低にそろえると，RMSEの値が極端に高いものや低いものになり，比較検討するのが難しくなることが考えられるためである。

比較する受験者数は，500，1000，2000，5000の 4 つとした。これらのそれぞれについて，共通項目数が2,4,6,8,10,12の 6 つのテストデータと，共通項目数が半分である場合の受験者の能力パラメータについて調べる。そのため，計24のRMSEを検討することになる。

Ⅱ-5　受験者集団の能力値差に着目した分析の分析手順

Ⅱ- 2 節の手順 1 では，受験者数を統一するため，無作為に受験者を抽出していた。ここでは受験者集団で能力値差がある場合を想定するため，受験者をある能力パラメータの範囲で抽出することにする。このため，元々のデータについて項目応答理論により分析し，受験者の能力パラメータを求めておく。

次に，受験者の抽出方法を 3 つに分け，テストデータを作成する。受験者集団の能力値差が「小さい」,「中程度」,「大きい」の 3 種類のテストデータを作成する。

能力値差小：両方の年度で受験者を無作為に抽出する。

能力値差中：一方の年度はθが 2 以下，もう一方の年度はθが -2 以上の受験者を無作為に抽出する。

能力値差大：一方の年度はθが 1 以下，もう一方の年度はθが -1 以上の受験者を無作為に抽出する。

これらのそれぞれについて，共通項目数が2,4,6,8,10,12の 6 つのテストデータと，基準テストの能力パラメータについて調べる。ここでは，計18のRMSEを比較することになる。また，受験者数は2000に，共通項目の項目弁別力については，Ⅱ- 4と同様に，中程度のものに統一した。

Ⅲ　結果と考察

Ⅲ-1　共通項目の項目弁別力に着目した分析

この分析は，共通項目に残す項目の項目弁別力の高さについて，高，中，低の 3 つに分けたうえで共通項目数の検討を加えたものである。実際に共通項目がどの程度の項目弁別力であったのかを，表3.1に示す。

表3.1 各教科の共通項目の項目弁別力

	英	国	数
共通項目 01	0.206	0.222	0.541
共通項目 02	0.237	0.223	0.575
共通項目 03	0.238	0.283	0.591
共通項目 04	0.363	0.312	0.597
共通項目 05	0.371	0.352	0.652
共通項目 06	0.428	0.362	0.673
共通項目 07	0.448	0.376	0.720
共通項目 08	0.468	0.423	0.760
共通項目 09	0.480	0.482	0.857
共通項目 10	0.546	0.516	0.863
共通項目 11	0.584	0.617	1.107
共通項目 12	0.587	0.651	1.177
共通項目 13	0.591	0.787	1.296
共通項目 14	0.632	0.869	−
共通項目 15	0.646	1.140	−
共通項目 16	0.724	−	−
共通項目 17	0.726	−	−
共通項目 18	0.754	−	−
共通項目 19	0.793	−	−
共通項目 20	0.870	−	−

さらに，共通項目の項目弁別力に着目した分析で得られた英語と国語のRMSEの値について，表3.2，表3.3と図3.1，図3.2にまとめた。

表3.2 英語における弁別力別のRMSE

弁別力	2	4	6	8	10	12
高	0.0154	0.0151	0.0131	0.0155	0.0174	0.0148
中	0.0310	0.0303	0.0175	0.0171	0.0119	0.0118
低	0.0587	0.0577	0.0534	0.0476	0.0405	0.0255

表3.3 国語における弁別力別のRMSE

弁別力	2	4	6	8	10	12
高	0.0172	0.0174	0.0142	0.0154	0.0090	0.0060
中	0.0269	0.0200	0.0235	0.0173	0.0115	0.0102
低	0.0174	0.0225	0.0146	0.0187	0.0152	0.0090

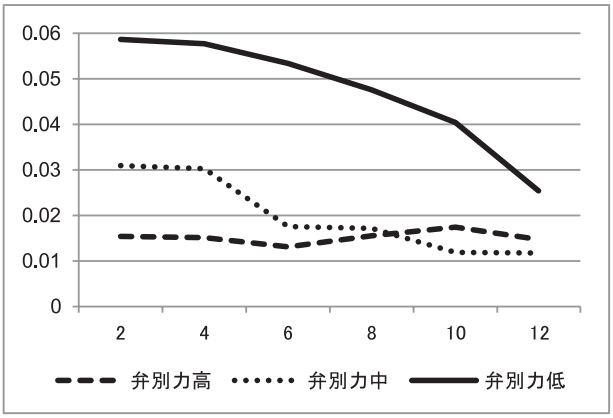

図3.1 英語における弁別力別のRMSE

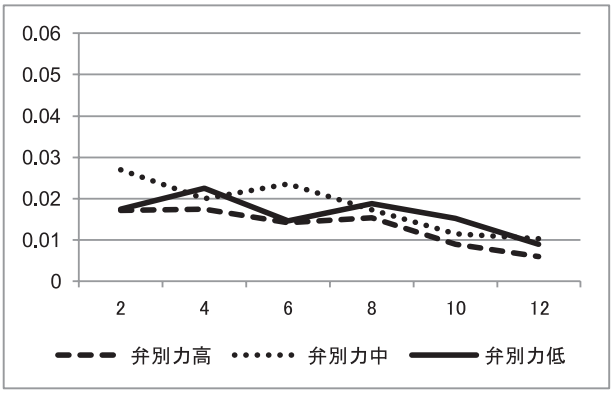

図3.2 国語における弁別力別のRMSE

共通項目の項目弁別力に着目した分析に対する考察

分析の結果，等化の精度の指標としたRMSEが得られた。これは，基準テストから得られた能力パラメータに対して，基準テストよりも共通項目を減らしたテストデータから得られた能力パラメータが，平均的にどの程度ずれているかを表す。

表3.2，表3.3，より，英語のRMSEの一番大きい値が0.0587であるのに対し，国語の場合は0.0270となっていることから，英語と国語を比べると，国語の方が推定の精度がよい。これは，全体的に国語における共通項目の弁別力が英語に比べて高いためであると考えられる。

全ての教科について，共通項目数が多くなると，弁別力の高低にかかわらず，RMSEの値が小さくなる傾向がみられた。つまり，共通項目数が増えると，等化の精度がよくなる傾向があるということである。

英語の項目弁別力が中程度である場合，図3.1，共通項目数が 4 から 6 にかけて大幅なRMSEの減少がみられることがわかる。この分析においては，共通項目数の目安として 6 以上あるとよいと考えられる。

また英語では，項目弁別力が高いほど，等化の精度がよくなる傾向がみられる。しかし，国語では，図3.2より，部分的に弁別力が高いテストデータのRMSEが低い方のRMSEを上回るところも見られる。また，弁別力が変わっても，RMSEは大きく変化しなかった。項目弁別力の高い項目を共通項目にしたとしても，等化の精度を大きく改善できない場合があると考えられる。

次に 2 教科のうち，弁別力の違いにより，RMSEに特に差が見られた英語に注目する。項目弁別力の低い場合と高い場合で，最もRMSEの差が大きかったのは，共通項目数が 2 の場合である。この差は，0.0433となる。今回の分析では，共通項目の項目弁別力の低いものから高いものに変えた場合，RMSEでいうと最大0.04程度等化の精度を改善している。

Ⅲ-2　受験者数に着目した分析

この分析は，受験者数が，500，1000，2000，5000のそれぞれの場合における等化の精度について着目したものである。共通項目に残す項目の項目弁別力は，共通項目の項目弁別力に着目した分析における中程度のものに統一している。受験者数に着目した分析で得られたRMSEの値について，表3.4，表3.5と図3.3，図3.4にまとめた。

表3.4 英語における受験者数別のRMSE

被験者数	2	4	6	8	10	12
500	0.082	0.078	0.060	0.024	0.026	0.020
1000	0.045	0.032	0.022	0.020	0.022	0.015
2000	0.031	0.030	0.018	0.017	0.012	0.012
5000	0.028	0.019	0.022	0.008	0.007	0.004

表3.5 国語における受験者数別のRMSE

被験者数	2	4	6	8	10	12
500	0.082	0.078	0.060	0.024	0.026	0.020
1000	0.045	0.032	0.022	0.020	0.022	0.015
2000	0.031	0.030	0.018	0.017	0.012	0.012
5000	0.028	0.019	0.022	0.008	0.007	0.004

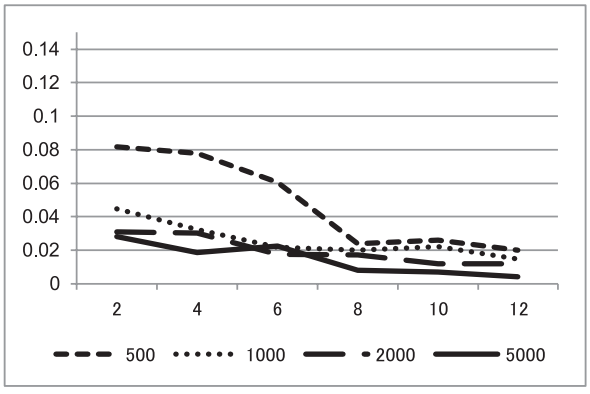

図3.3 英語における受験者数別のRMSE

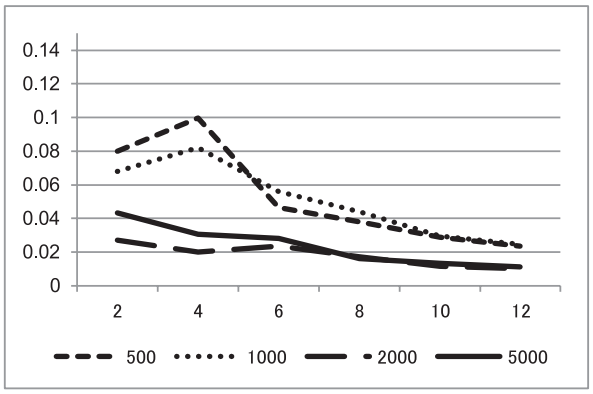

図3.4 国語における受験者数別のRMSE

受験者数に着目した分析に対する考察

共通項目の項目弁別力に着目した分析と同様に，図3.3と図3.4をみると，共通項目数が多くなるほど，RMSEの値が小さくなる，つまり等化の精度が高くなる傾向が見られた。

英語では図3.3より，共通項目数が 6 から 8 であるときにかけて，受験者数が500のテストデータのRMSEが大幅に減少している。また，国語では，図3.4より，共通項目数が 4 から 6 であるときにかけて，受験者数が500のテストデータのRMSEが大幅に減少している。これらのことを踏まえると，この分析において，共通項目数の目安として， 6 から 8 以上が望ましいと考えられる。

英語と国語に関して，受験者数が少なくなるほど，等化の精度が悪くなる傾向が見られた。ただし，受験者数が2000と5000の場合，RMSEの値が逆転しているところも見られる。ここから，受験者数は2000程度であれば，等化の精度という観点では十分な数とみなすことができると考えられる。

また，英語では受験者数が500の場合，国語では500, 1000の場合，共通項目数が 4 以下の場合，特にRMSEの値が高くなっていることが分かる。受験者数が2000を下回り，共通項目数が 4 以下の場合，等化の精度に注意する必要があると考えられる。

英語において，受験者数が500と5000の間で最もRMSEの差が大きかったのは，共通項目数が 4 の場合である。この差は，0.0594となる。同様に，国語では共通項目が 4 の場合に，受験者数が500と5000の間のRMSEの差が0.0691となっている。

Ⅲ-3　受験者集団の能力値差に着目した分析

この分析は，受験者集団の能力値差が小さい場合，大きい場合，中程度の場合について着目したものである。このため，各教科の受験者の能力値をあらかじめ分析し，能力値差を設けた場合のテストデータを作成する必要があった。この分析において，共通項目に残す項目の項目弁別力は，共通項目の項目弁別力に着目した分析における中程度のものに統一している。また，受験者数は2000に統一している。

ここで，受験者集団の能力値差がどの程度生まれたのかを示すため，表3.6に各教科のそれぞれの受験者集団の平均（標準偏差）とその差を示す。

表3.6 各テストデータの受験者の能力パラメータθの平均(SD)と平均値差

英語			
	θ低群平均 (SD)	θ高群平均 (SD)	θ平均値差
能力値差大	0.009 (0.865)	1.854 (0.649)	1.845
能力値差中	−0.002 (0.914)	0.078 (0.892)	0.080
能力値差小	−0.01 (0.902)	0.031 (0.855)	0.041
国語			
	θ低群平均 (SD)	θ高群平均 (SD)	θ平均値差
能力値差大	0.010 (0.873)	0.537 (0.844)	0.527
能力値差中	0.013 (0.887)	0.126 (0.891)	0.113
能力値差小	0.011 (0.887)	0.024 (0.902)	0.013

受験者集団の能力値差に着目した分析で得られたRMSEの値については，表3.7，表3.8と図3.5，図3.6にまとめた。

表3.7 英語における能力値差別のRMSE

能力値差	2	4	6	8	10	12
大	0.3579	0.0963	0.0733	0.0600	0.0384	0.0400
中	0.1132	0.0348	0.0259	0.0208	0.0194	0.0222
小	0.0310	0.0303	0.0175	0.0171	0.0119	0.0118

表3.8 国語における能力値差別のRMSE

能力値差	2	4	6	8	10	12
大	0.2740	0.1524	0.1081	0.0599	0.0431	0.0212
中	0.0871	0.0644	0.0423	0.0154	0.0151	0.0086
小	0.0269	0.0200	0.0235	0.0173	0.0115	0.0102

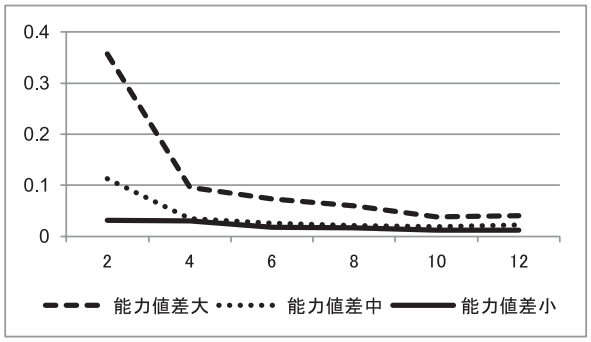

図3.5 英語における能力値差別のRMSE

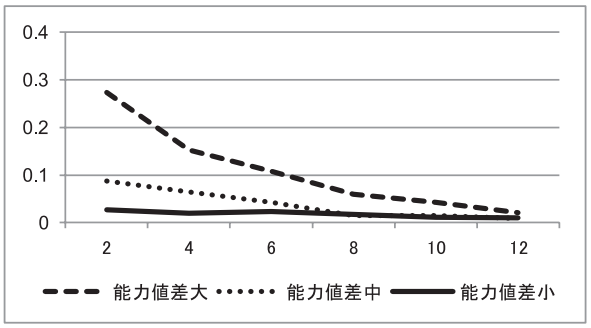

図3.6 国語における能力値差別のRMSE

受験者集団の能力値差に着目した分析に対する考察

表3.7，表3.8より，共通項目の項目弁別力に着目した分析，受験者数に着目した分析と比べると，大きな値になっていることが分かる。特に，英語の能力値差が大きいテストデータの共通項目数が 2 の場合のRMSEは0.3579と，本研究中で最も大きい値となった。また，国語においても，共通項目数が 2 の場合，RMSEが0.2740と，英語についで大きな値を示している。

英語と国語に関して，受験者集団間の能力値差が大きいほど，また，共通項目数が少ないほど等化の精度が悪くなる傾向が図3.10，図3.11から読み取れる。特に，共通項目数が 2 である時，急激にRMSEの値が高くなっている。英語の場合，共通項目が 2 の時の，能力値差が小さい場合と大きい場合のRMSEの差が，0.3269となっている。同様に，国語の場合は0.2470となった。

英語では図3.10より，能力値差が大，中のテストデータについて，共通項目数が 2 から 4 であるときにかけて，受験者数が500のテストデータのRMSEが大幅に減少している。しかし，能力値差が大きいテストデータの場合，RMSEの値は共通項目数が 4 の場合でも0.1程度となっている。共通項目の項目弁別力に着目した分析，受験者数に着目した分析のRMSEの値と比べると，RMSEの値は未だ高いといえる。RMSEが0.1よりも小さくなることを等化の精度の基準として考えると，英語では共通項目数が 4 から 6 以上，国語では共通項目数が，8 以上が望ましいということになる。これらのことを踏まえると，この分析においては，共通項目数の目安として，6 から 8 以上が望ましいと考えられる。

Ⅳ 全体考察

本研究の分析全体を通じて，共通項目数が多いほど等化の精度がよくなるという傾向を確認することができた。ただし，英語，国語の場合，共通項目数が 8 項目よりも多い場合，等化の精度は大きく向上していない。特に等化の精度が大きく変動したのは，共通項目数が 2 から 6 の間の場合であった。

共通項目数を増やすことによって，等化の精度をよくすることが効果的に働くのは，共通項目数が特に 6 以下の場合であるといえる。

ただし，今回の分析にあたって，元のテストデータにあった項目を一部削除することで，非共通項目を作っている。そのため，今回の分析は総項目数の多くない場合であるといえる。

項目弁別力，受験者数，受験者集団間の能力値差が等化の精度に及ぼす影響について比較検討する。

本研究において，項目弁別力が等化の精度に大幅な影響を及ぼすについては，大きなものは見られなかった。英語では，共通項目の項目弁別力の低いテストデータと高いテストデータで，最大0.04程度のRMSEの差が確認できた。しかし，国語では大きな差が確認されなかった。

一方，受験者数が等化の精度に及ぼす影響については，英語，国語共にRMSEの差を確認することができた。受験者数が500と5000の間で，英語では0.06程度，国語では0.07程度の差が最大で見られる。ただし，受験者数が2000を超える場合，英語と国語ではRMSEに大きな差が見られない。

項目弁別力，受験者数と比較すると，受験者集団間に能力値差がある場合，等化の精度が大幅に悪化することが確認できた。能力値差が大きく，共通項目数が 2 の場合，RMSEの値は英語で0.3579，国語では0.2740と，本研究で特に高い値を示した。能力値差が低い場合と比較すると，英語では0.3269，国語では0.2471と，RMSEの差という観点からも大きな値を示した。

共通項目の項目弁別力が低い場合や受験者数が少ない場合よりも，受験者集団間に能力値差がある場合，特に等化の精度が悪くなるということが考えられる。このことから，垂直的等化を行う場合や，テストを長期的に経年比較する場合など，受験者集団の能力値が集団間で異なる場合，等化の精度に注意する必要があると考えられる。

本研究では，英語や国語の他に，数学の分析も行っていた。しかし，ソフトウェアで解が収束せず，適切なパラメータの推定値を求めることができなかった項目があったことと，項目弁別力の推定値が極端に高い項目があっ

たことから，数学のテストデータを分析に含めることができなかった。項目弁別力の推定値が極端に高い項目は，今回の数学の問題で局所独立の仮定を満たしていないと考えられる項目であった。項目応答理論による分析を行う場合，一次元性や局所独立の仮定について確認するだけでなく，分析した後に，どの項目が項目応答理論に適しているのかについて十分吟味することが重要であると考えられる。

Ⅴ　今後の課題

本研究では，分析で扱ったテストデータは，各教科で，1 つのテストだけであった。他のテストデータを分析することができれば，今回の分析で得られた結果の確認や，比較検討を行うことができる。また，今回は数学のテストデータについて，適切な推定値を得ることができなかったが，他の数学のテストデータと比較することで，このことが今回扱ったテスト問題によるたまたまの結果なのか，数学という教科の特性によるものなのかといった検討を加えることができるだろう。

今回の研究では，共通項目の項目困難度が等化の精度に及ぼす影響について触れることができなかった。豊田(2002)[6]は共通項目として，困難度母数の値の違いの大きい項目が望ましいとしている。今回の研究で，受験者集団の能力値差が大きい場合に等化の精度が悪くなっていた。ここで，共通項目の困難度にばらつきを持たせることで，どの程度等化の精度が改善することができるのかについて，検討を加えることができる。

本研究では，共通項目数，共通項目の項目弁別力，受験者数，受験者集団間の能力値差によって，等化の精度がどのように変化するのかについて分析を行った。この他に，推定の方法，等化の方法，モデルの選択，テストデータのモデルの適合度など，等化の精度に影響すると考えられる要素は多岐にわたる。例えば，本研究では2PLMによる分析を行ったが，3PLMによる分析を行えば，共通項目の当て推量パラメータが等化の精度に与える影響について検討することもできる。さらに，今回の結果は本研究で用いたテストの受験者の能力パラメータから算出したものである。本研究で示した指針は，本研究で扱ったような一般的な高校生の基礎学力を測るテストにおいては有用であると考えられる。しかし、指針とする共通項目数は，相対的な受験者の能力パラメータによって変わってくる可能性があるため，この指針を一般化することはできない。

しかし，実データを用いた等化の精度の研究を探索的に進め，様々なテストデータや異なるモデルについて検討を加えることで，特定の状況だけでなく，多様なテスト開発の場面に合った有用な知見を得ることができるだろう。

ー謝　辞ー

本論文は平成22年度岡山大学教育学部卒業論文を加筆修正したものである。ベネッセコーポレーションの金森保智様，対馬英樹様，木内祐輔様をはじめ，研究にご協力いただきました皆様に心より感謝申し上げます。

ー文　献ー

(1)　孫媛「テスト得点の精度を吟味する古典的テスト理論」 渡部洋編『心理統計の技法』 福村出版, pp.99-112, 2002

(2)　Hambleton, R. K., Swaminathan, H., & Rogers, H. J. *Fundamentals of Item Response Theory.,* Newbury Park CA: Sage Press, 1991

(3)　大友賢二『一言語テスト・データの新しい分析法―項目応答理論入門』 大修館書店, 1996

(4)　Petersen, N. S., Kolen, M. J., & Hoover, H. D. In R. L. Linn (ed.), Scaling, norming, and equating., *Educational measurement* 3rd ed., New York American Council on Education and Macmillan., pp.221-262, 1989池田ほか編訳『教育測定学』第 3 版, みくに出版, 1992

(5)　藤森進「同時尺度調整法による垂直的等化のシミュレーションによる検討」岡山大学教育学部学術研究委員会『岡山大学教育学部研究集録』, 97, pp.173-177, 1997

(6)　豊田秀樹『項目応答理論[入門編]―テストと測定の科学―』 朝倉書店, 2002

(7)　藤森進「同時尺度調整法による垂直的等化の検討」文教大学人間科学部『人間科学研究』,20,34-47 1998

(8)　前川眞一・菊池賢一・内田照久・中畝菜穂子・石塚智一「大学入試センター試験得点の標準化の試み―項目応答理論による方法―」『大学入試研究ジャーナル』13, pp.81-87, 2003

(9)　熊谷龍一・山口大輔・小林万里子・別府正彦・脇田貴文・野口裕「大規模英語学力テストにおける年度間・年度内比較―大学受験生の英語学力の推移―」『日本テスト学会誌』3, pp.84-90, 2007

(10) De Ayala, R. J., Plake, B. S., & Impara, J. C. The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement,* Vol.38, pp.213-234, 2001