クラスルームテストの作成・評価方法の検討：多枝選択式項目作成ガイドラインと識別力指標について

名古屋大学大学院教育発達科学研究科

坪田　彩乃

目　次

第1章
本論文の背景と目的

1.1 はじめに

「三人の学生」というタイトルの短編物語がある。これはシャーロックホームズシリーズのひとつの話である。物語の内容は，奨学金を得るための試験の前日，テストの作成者が印刷された試験問題を机上に置いたまま部屋を離れた際に，何者かがその試験問題を盗み見たことが明らかとなり，それが可能であった三人の学生のうち誰が実行したのかをシャーロックホームズが突き止めるというものである。ここでは，試験問題を盗み見た者に利が大きいことから，試験の公平性が損なわれるということが問題となった。明確な事件ではない謎を解決するという点で，この短編はシリーズの中でも異色の存在だという。

物語上でも話題の中心になるように，テストというものへは強い関心が寄せられる。その関心の焦点は様々で，そのテストが公平に行われるかどうかであることもあれば，そのテスト自体が理に適ったものであるかどうかということもある。

さて，「三人の学生」に登場する大学については，オックスフォード大学であるとかケンブリッジ大学であるとか，意見が分かれるところだという。しかし，ここで明白なのは，この試験が口述試験ではなく，紙と鉛筆を利用した筆記試験であったということである。

古くからヨーロッパでは，口述試験が主流であった。しかし口述試験では，評価の公平さや困難度の設定の難しさなどが問題になる(印東・牧田・肥田野, 1950)。こうした背景により，客観的に行うことが可能な筆記試験が望ましいとされた。十八世紀初めには，ケンブリッジ大学で筆記試験が行われるようになったといわれている。一方東洋では，科挙に代表されるように，古くから筆記試験による評価が行われてきた。

昨今では，入学試験や資格試験，クラスルームで行われるテストなど幅広いテストで筆記試験が用いられている。初期の頃は論述式テストが主流であった筆記試験も，時代と共に新たな項目形式が用いられるようになった。現在では，多くの試験で多枝選択式による項目をはじめとする客観式テストが用いられている。

時代によってテストは絶えず変化してきた。本研究は，現在日本におけるテスト文化の中で行われているテスト，特にクラスルームで行われているテストに焦点をあて，テストの作成・評価に関する知見を提供することを目指す。

1.2 クラスルームテストとは

クラスルームにおける教育評価

教育評価とは，教育実践自体への反省から修正や改善をすることを目的とするものである。そしてそこで行われる評価は，教育目標を評価基準とした「目標に準拠した評価」である。

学力に関しての「目標に準拠した評価」は，基礎学力に加え，発展的な高次の学力育成を目指すとともに，学びと学力形成の実態をより確かな形で捉えようとするものである(若林, 2021)。

ブルームは授業過程で実施される学力評価を，「診断的評価」「形成的評価」「総括的評価」の三段階に分化した。それぞれの段階において有効なフィードバックを行うことで，児童・生徒の学力や発達を保障することになるという(田中, 2008)。

診断的評価とは，新しい課程，学年，授業，単元に入る前などに，指導の参考にするための事前情報を収集することを目的として行われる評価である。

形成的評価とは, ある単元の学習を進める際に, 学習状況を確認し, つまずきの早期発見・早期回復によって児童・生徒の学力形成に利用することを目的として行われる評価である。

総括的評価とは，課程，学年，学期，単元が終了する時点での広範囲にわたる学力の成果をまとめ，成績づけに利用するための評価である。

これらの各段階で学力を測定するために，クラスルームではテストが用いられる。つまり，クラスルームにおけるテストとは，児童・生徒の学力を測定し，その結果を教育にフィードバックするまでを目的として行われる。

クラスルームテストの作成・使用

クラスルームによるテストでは通常，教師が作成・実施・採点という一連の流れを行う。そして, 実施されたテストの採点結果は児童・生徒の成績評価の際の資料として用いられる。

一般的に，テストという測定道具を作成・実施する際には，利用目的や場面に応じた基本設計が構想され，適切に開発されたテストを適切に実施することが求められる(日本テスト学会, 2007)。テストに関わる上で，どのようにテストを扱っていくかという点は非常に重要であり，それらについて共有される必要がある。一方で，教員養成課程において，テストの作成や実施という評価に関するカリキュラムに基づいた教育は行われていない(若林・根岸, 1993)。2019 年に改訂された教職課程科目に関しても，学習評価に関わる記載は教育方法論の科目内における到達目標のひとつに過ぎず，その具体性への言及はなされていない。しかしながら一般的に，教師はその職に就くと同時に，担当科目に関するテストを作成することが前提となる。

クラスルームでの児童・生徒の評価方法については, 国立教育政策研究所が提示する「『指

導と評価の一体化』のための学習評価に関する参考資料」などが存在するが，テスト作成に関しての言及はない。学力テストを作成するために参考となる書物は，たとえば矢口(1957)，橋本(1981)などが挙げられる。矢口(1957)は，国語科，社会科，算数・数学科，理科の科目別に，小学校・中学校でのテストに分け，単元別にどのようなテストを作成することが望ましいかを解説している。橋本(1981)は，教師が作成する到達度評価や，標準化された到達度評価を行うためのテストの作成方法について，実施方法から項目の選定，テストの評価方法に至るまでの解説をおこなっている。また，言語テストや英語科のテスト作成に関する書籍は他にも存在している。しかしながら，科目や単元によらないテストの作成や評価の方法，たとえば児童・生徒の能力を引き出すような項目作成の方法や，実施したテストへの評価に対する関心は決して高いとはいえない。これらの知見が共有されていない状態では，テストの作成・実施について個人の経験や勘に基づいて作成することとなる。

クラスルームテストと指導要録

指導要録とは，指導機能と証明機能という二つの機能をもつ法的に義務づけられた資料であり，児童・生徒が転校する際に転校先に複写を渡すという使い方や，進学や就職の際に提出する内申書の原簿ともなる(樋口, 2021)。

2001(平成 13)年改訂指導要録では，評定の位置づけが目標に準拠した評価に改められ，「わかる・できる」という具体的内容の到達を表す規準による評価が求められるようになり(八木, 2006)，それまでの相対評価(段階評定の各段階がクラスで何名と人数が決まっている評価)からは大きな変更がもたらされた。相対評価とは，集団に実施したテスト結果に基づき，あらかじめ統計的に設定された基準に照らして解釈するという評価法である(橋本, 1979)。相対評価をクラスルームに適用することについては，①非教育的な評価論である，②排他的な競争を常態化させる，③学力の実態を反映しない，④教育活動を評価できないという問題が指摘されていた(田中, 2002)。

「目標に準拠した評価」では，教育目標がどの程度到達できるかどうかを測ることができるテストの作成が求められる。先述したように，クラスルームテストは教育評価において学力を測定する資料のひとつとして活用される。つまりクラスルームテストの資料は，場合によっては進学先の選定材料として用いられるなど，児童・生徒の処遇へと影響を及ぼすものである。

教育現場における児童・生徒の評定方法が「目標に準拠した評価」を前提としたものへと切り替わったとしても，こうした評定の一端を担うテストが十分な機能を果たしていないとき，様々な問題が生じる。そのため次節では，テストにどのような諸問題が生じるかについて論じていく。

1.3 テストにまつわる諸問題

テストとは，教育成果の実態について目に見える形での情報を与えてくれる唯一の道具である(池田,2006)。この道具が適切でない場合，様々な問題が生じる。たとえば，受検者の能力を引き出すのに十分ではない試験問題であったとき，受検者の能力の実際が反映されない評価をつける可能性がある(池田, 1992)。

これらを踏まえ，本節ではテストに関する問題を，測定道具としてのテストに関する問題，テスト項目の作成に関する問題，テストの評価に関する問題という三つの側面から取り上げる。

測定道具としてのテストに関する問題

初めに，測定道具としてのテストに関する問題について取り上げる。

テストは，その扱い方に着目されることが多い。入学者選抜試験など受検者の処遇に関わるようなハイステークスな場面では，特にそれが顕著に表れる。たとえば，採点ミスによって合否に影響を及ぼすことは，受検者の一生に影響を及ぼすことに繋がる。こうした事態が生じたとき，主要メディアによって大々的に報道され，多くの関心が寄せられる。

本来，テストは正しく扱われれば，特に学校場面においては教育目標に対しての有用な資料となりうる。しかし，テストそのものの出来が悪かったり，テストから得られた結果についての扱われ方が適切ではなかったりするとき，児童・生徒にとって悪影響を及ぼすことも生じる。

テストから得られた評価が実態を反映しないとき，その評価に合うように実態が変わることがある。たとえば，教師自身の期待は生徒の学力へ影響を及ぼす要因になりうる(東, 2001)。ピグマリオン効果(Rosenthal & Jacobson, 1968)やラベリング理論など，教師の期待や先入観によって児童・生徒の行動や学力へ影響を及ぼすことがある。こうした教師の視点の背景に，クラスルームで行われるテストが一切介在しないという保証はない。また，テストの評価そのものは，児童・生徒のテストへの動機づけにも影響を及ぼす。梶谷・小林・鈴木・中田・盛本(2013)は，学生に成績順位を通知した際に，成績上位者と成績下位者では次の試験に対しての行動が異なっていたという。特に，上位の成績と通知された学生は，次回の試験に対して油断するという油断効果が見られた。つまり，試験の成績そのものよりも，自分が相対的にどの群に入るかどうかによって，次の試験への動機づけが異なるという。

教育現場で行うテストでは，受検者個人の成績の参照が可能である。そのため，不適切な評価により，本来の能力を反映しないような教育上の判断をされた場合，それによって不利益をもたらす可能性も否めない。つまり受検者の成績評価は，クラスルームテストにおいて教育指導場面にまで及ぶと考えられる。しかしながら，テストの扱い方がいくら適切であったからとはいえ，そもそもそのテスト自体が学力の測定道具として意味を成さないもので

あれば，導かれる結論は実態を反映しないものとなる。

テスト項目の作成に関する問題

テストは教育測定を行うための道具である。そのため，テストに含まれる項目は，受検者の能力(学力)を測定するに値する項目である必要がある。たとえば，授業で行った教育の効果を測定するためのテストにおいて，授業で扱わなかった単元の知識を要する項目は不適切であると考えられる。授業による効果を十分に測定しうるような項目を作成するために，出題範囲が適切であるかどうかを十分に吟味する必要がある。未学習の単元が前提となる項目に正答できたことは，授業の効果を反映しているものではない。せいぜい先取り学習がどの程度進んでいるかの判断材料になるかもしれないが，それを問うべきはテスト以外の手段を用いるほうが相応しいだろう。

テストの項目が測定したい範囲を反映されていると考えられる場合でも，項目の機能が十分に働かないことがある。多枝選択式のテスト項目に欠点 (Flaw) があることでテストが測定道具としての機能を十分に果たさないことがある。項目の Flaw が正答率や識別力へ影響を及ぼすことは Downing (2005) や Martínez, Moreno, Martín, & Trigo (2009) で確認されているものの，これらはテストの規模によらず存在しており，Flaw の一切ない項目からのみ成るテストを作成することは容易ではない。Rush, Rankin, & White (2016) は，学内で用いられた試験問題 1,925 項目のうち 37.3%の項目で 1 つ以上の欠点が見つかったと報告している。また，Tarrant, Knierim, Hayes, & Ware (2006) でも同様に，看護師試験で用いられる多枝選択式項目の 46.2%の項目で 1 つ以上の欠点が見つかった。また，項目に Flaw があることで，テストワイズネスにより受検者へ正答の手掛かりを与えるものもある (Chittooran, & Miles, 2001)。

このように，テストが問いたい内容以外の側面でも，テストの機能を低下させる要因が存在する。こうした要因は，テスト作成の時点で十分に留意されるべき点であるが，Flaw が一つもないテストを作成することも現実的には難しい。そのため，できる限り Flaw の影響を小さくするような項目作成をする姿勢が求められる。

テストの評価に関する問題

テストを作成する段階で，その内容や細かな点に至るまで十分に吟味されたとしても，実際に作成者の意図に沿った形で機能していたかを確認する必要がある。たとえば，出題ミスがなかったか，テストの内容が目的に即したものであったか，受検者集団に沿ったテストとなっていたかどうかをはじめ，テストの正答率や識別力が適切であったかどうかをも検討することが求められる。そして，事前に検討した通りに機能していないとき，具体的にどの項目が機能していなかったのか，なぜ機能しない項目であったのかを明らかにし，評価の対象から外す項目とするのかどうかを検討したり，次にテスト項目を作成する際に注意する

べき点として記録に残したりする必要がある。

池田(1982)は，テストを評価することで明らかとなる問題のひとつに，天井効果，床面効果を挙げている。これは，テストが受検者集団に対して易しすぎたり，難しすぎたりすることで，測定道具としての機能を損なっている状態をいう。受検者集団に対して適切な難易度のテストでないとき，受検者の得点分布は歪む。こうしたテストは，項目が全体的に易しいことで多くの受検者が高得点となったり(天井効果)，反対に，項目の難易度が高すぎて多くの受検者が低得点となったり(床面効果)する。これらの問題は，そのテストにより難しい／易しい項目が含まれていれば生じなかった問題である。

テストには，全国的に一斉に行うようなもの(e.g.大学入学共通テスト，全国学力・学習状況調査)や，各教員が個別に行うクラスルームテストがある(日本テスト学会, 2007)。これらは，テストの目的，受検者集団や作題者など，様々な点で異なり，それに付随してテストの評価という点においても大きく異なる。

大規模テストは，テスト項目の作成，実施，採点，評価に至るまで複数の人員の手で進められる。そして，テストの評価はテストの専門家が行っている。中には，古典的な手法である項目分析のみならず，項目応答理論などを用いて評価を行う。これらを通して，テストが適切なものであったかを確認し，また，テストの改善にも役立てられる。

一方で，クラスルームで行われるようなテストでは，授業実施者がテストの作成・実施をし，児童・生徒という受検者にフィードバックを与えるまでの一連の作業を行う。しかし，クラスルームで行われるテストの評価はほとんどされていない。

テストの規模によらず，実施されたテストは評価されることで，そのテストが目的に適ったテストであったか，そこから得られるテストの得点は意味を成すのかを検討することができるようになり，更にその後のテスト作成に役立てることができる。しかしながら，クラスルームで行われるテストでは評価が十分になされていない。テストの評価がなされていないとき，そのテストで得られた結果の信用が揺らぐ。テストを十分に活用するためにも，テストは評価され，その機能について確かめられる必要がある。

このように，テストに関するあらゆる問題は，最終的には受検者の評価・処遇へと影響を及ぼす。そのため，テストを活用する際には，それぞれの問題について理解した上で影響を最小限に留めるように努めたい。また，こうした問題が生じることを踏まえ，どのようにそれらに対処することが可能かについても検討されることが求められる。

1.4 本論文の目的と構成

本章で述べたように，テストの作成・実施・評価について体系的に学ぶ機会は少ないにも関わらず，日々様々な場面でテストが作成され，実施されている。テストの作成に対して無頓着であったり，実施したテストについて顧みたりしなければ，本来テストから得られたであろう結果を十分に手に入れられなかったり，得られた情報の質が低下したりする。そして，テストは時に受検者の処遇に大きな影響を及ぼす。テスト実施者が求めている情報を効果的に得るためにも，よいテストを実施することが必要である。

そのため本論文では，よいテストを作成するための方法として，テスト項目の作成時ならびにテスト実施後の評価にかかわる問題について検討を行う。これらの研究を通し，テスト研究者をはじめとするテストの専門家のみならず，教師，テスト事業者，社内人事の担当者などの教育現場や職場などでテストに触れているあらゆる非専門家に向けて，テストの作成・実施・分析を行う際に有用な知見を示すことを目的としている。

第 2 章において，本論文で取り上げる研究の理論的背景について述べる。最初に「よいテスト」とはどのようなテストであるかを検討する。その上で，それを達成するためにテストの作成時に確認したい項目作成ガイドラインとはどのようなものであるのか，テストの評価時に検討するべき項目分析，特に識別力について説明する。

第 2 章での検討を踏まえ，第 3 章では，テストの作成に関する研究を行う(研究 1)。特に多枝選択式テストを作成する際に有用となる多枝選択式項目作成ガイドライン(Haladyna & Rodriguez, 2013)について, 作題者が特に重視するべきガイドラインを確認する。これにより，より受検者の能力を反映するテスト項目の作成方法への示唆を与える。

第 4 章では，テストの評価に関する研究を行う(研究 2)。ここでは，テストの非専門家でも扱いやすいとされるテストの識別力指標である D 指標について取り上げる。D 指標を求めるために有用となる群分けの基準，ならびにそれを達成するためのテスト条件について提案する。これにより，クラスルームテストの評価を行う手法に関しての示唆を与える。

これらを踏まえ第 5 章では，テストの作成・実施・評価時にテスト作成者が注意するべき点について，総合的に考察する。本研究で明らかとなったテスト作成時に注意すべき点や，テストの評価方法を踏まえ，具体的にどのようなテストを作成することが望ましいかについて論じる。また，今後の展望についても述べる。

図 1-1 は，本論文の章立ての関連性を示すものである。本章では，教育場面におけるテストの役割について論じてきた。続く第 2 章では，テスト研究の文脈を踏まえ，テストに生じる問題の解決方策について検討をする。ここでの検討を踏まえ，第 3 章では，テスト作成時における問題解決策について実証的に検討し，第 4 章ではテスト評価時における具体的な方策について検討を行う。第 5 章では，それらの研究を踏まえ総合的に考察する。

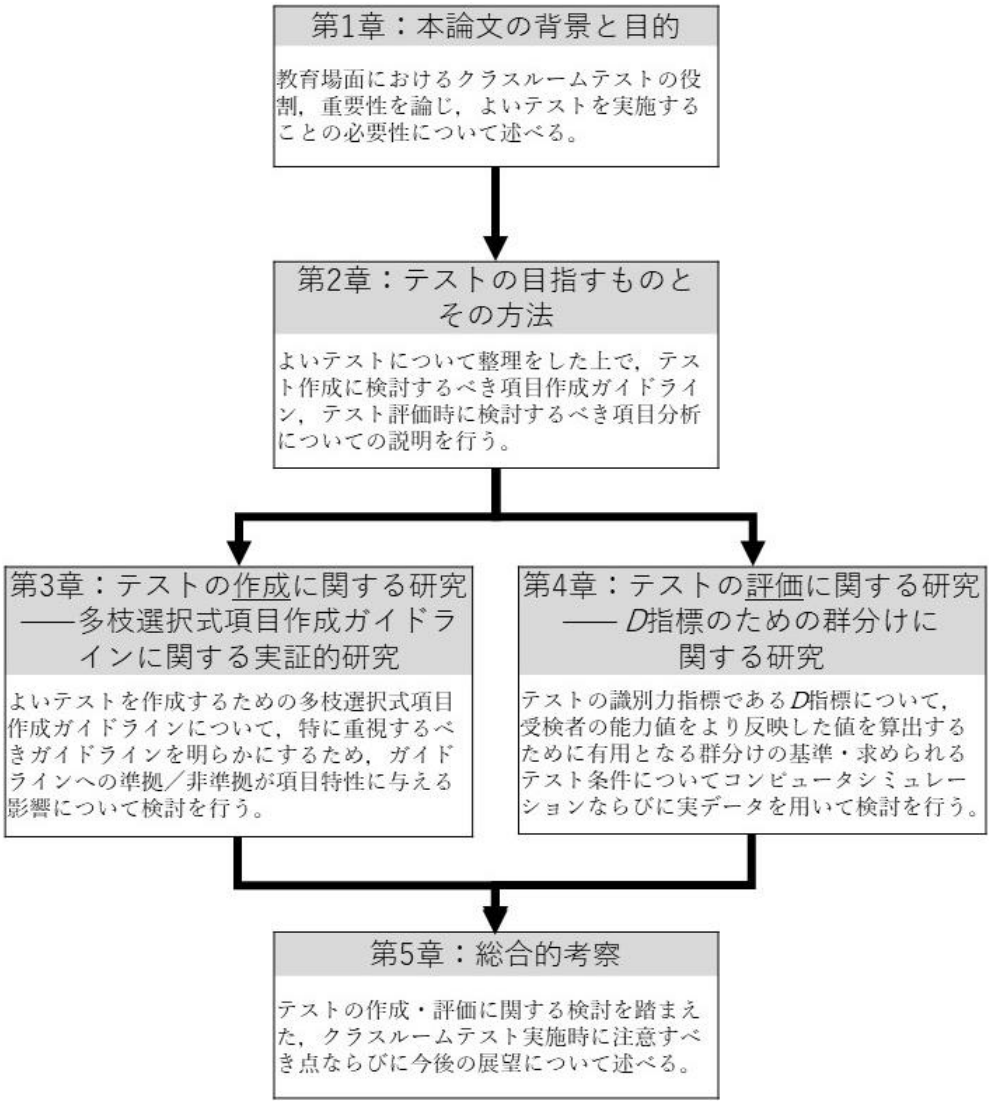

図 1-1 本論文の構成[1]

[1] 関連業績

第 3 章

坪田彩乃・石井秀宗(2019). 受検者は項目の flaw に気が付くのか―項目作成ガイドラインの実験的検討― 日本テスト学会第 17 回大会，名古屋大学

坪田彩乃・石井秀宗(2020). 多枝選択式問題作成ガイドラインの実証的検討　日本テスト学会誌,16,1-12（査読あり）

Tsubota.A & Ishii.H(2021). What item flaws affect item characteristics without recognized by examinees?　International Congress of Psychology. online

第 4 章

坪田彩乃・石井秀宗(印刷中).　D 指標を用いた項目分析のためテスト条件の検討 ――より真値に近い状態を目指して Quality Education（査読あり）

第2章
テストの目指すものとその方法

2.1 テストの目指すものとは

テストは受検者の学力を測定する道具である。そしてテストが目指すべきものは，測定道具として有用であるもの，すなわち「よいテスト」となることである。

日本テスト学会(2007)は，よいテストの作成・実施・利用・管理に際して，テストの様々な条件を考慮し，よいテストとはどのようなものであるかを考え，責任をもって適切に対処することが重要であるとしている。

よいテストについて，日本テスト学会(2010)は，明確な目的の上で，その目的を果たすように作成・実施・採点・運用されているテストであるとし，小泉(2017)は，信頼性・妥当性・実用性を兼ね備えたものであるともいう。

テストがどのような目的をもって行われるのか，どのような集団を対象として行われるのかなど，テストに関する様々な条件が異なれば，同じテストを用いても信頼性・妥当性・実用性は変化する。大規模な集団に対して行うのであれば適しているテストを，一部の能力層からのみ抽出した小さな集団で行ったときには，受検者の得点が一部に集中してしまうなど必ずしも適しているとはいえない。つまり，よいテストというものは絶対的に一意に決まるわけではない。

そこで本節では，テストの基本設計について紹介した後，テスト実施の順序に則り，テストの目的，項目作成，結果の利用に焦点をあて，よいテストを達成するために必要な観点を整理していく。

テストの基本設計について

日本テスト学会(2007)によると，テストを作成する際には利用目的や場面に合わせ，測定する内容，形式，実施方法や結果の利用方法といった基本設計を行うことが求められる。つまり，テストの設計を行う際には，それが何を目的としたテストであるのか，何を測定するテストであるか，そしてどのように結果を用いるのかを事前に検討することが望ましいとされている。これらの観点をテスト実施前に整理することにより，よりテストを効果的な測定道具とすることができるのである。

そしてテストの基本設計は，テストの作成・実施・評価という一連のプロセスにおける基準となる。テストの作成・実施は，この基本設計を踏まえて行われるし，テストの評価は，基本設計に則ったテストであったかを確認する。つまり，テストの目的に応じて吟味された基本設計は，項目の作成，結果の利用のいずれの段階においても重要となる。

テストの目的について

テストには，入学試験，学力試験，資格試験など，様々な種類のものが存在する。また，学力を測定するためのテストといっても，大規模に行われる学力調査であるのか，選抜を目

的とした試験であるのか，クラスルーム単位で行われるテストであるのかによって，その目的は異なる。

日本テスト学会(2010)は，グループを対象としたテストとして学力調査を，個人を対象としたテストとして選抜テストとクラスルームテストがあるとしている。また，それぞれの目的として，学力調査はグループの学力の全体像を具体的かつ広範に把握することを，選抜テストは受検者の合否が明確にわかることを，クラスルームテストはクラス内の学習状況を把握することを挙げている。

こうしたテストの目的に応じて，求める測定内容も異なる。学力調査では，全体の学力の程度を測定することを目的としている。このようなテストでは，受検者の個々の学力を測定することではなく，学力の全体像を把握することが目的となる。この場合，どこの学校に所属しているかなどといった受検者集団の属性によって結果への影響が生じないような内容とすることが求められる。すなわち，ある学校では習っていた単元が別の学校では習っていないということがないように，また，ある地域では馴染みが深いために正答することが容易であるが別の地域では馴染みがないために解答に労力を割くなどということがないように十分に検討した上で作成されることが求められる。

一方で，個人を対象とした選抜テストやクラスルームテストでは，個人個人の学力を測定することを目的としている。特にクラスルームテストでは，授業の学習状況を把握することを目的としており，授業で扱った単元について児童・生徒ひとりひとりの到達度を測定することを目的とするテストである。また，クラスルームテストでは，受検者ひとりひとりに焦点をあてた結果のフィードバックが行われ，各個人の学習への参考資料となることが求められる。これらのことより，扱われる内容は授業で取り扱った単元についての到達度を測定するのに相応しいものであり，誤答した項目からどこで躓いているのかを確認できるような設計であることが望ましい。

テスト項目の作成について

テストの目的や測定内容が十分に吟味されたら，その目的が達成されるようなテストを基本方針に照らしながら作成していく。つまり，実際のテスト項目を作成する。

テストで用いられる主な項目形式として，記述式項目と多枝選択式項目が挙げられる。記述式項目は受検者の筆記による解答を求める形式であり，短答式，穴埋め式，論述式などが含まれる。多枝選択式項目は複数の選択枝から正答を解答させるという形式である。これは設問部分の「幹」と選択枝の「枝」から成る(池田, 2007)。

これらの形式にはそれぞれメリットとデメリットが存在し，互いに補い合うよさがあることから，テストの目的によって柔軟に組み合わせることがよいとされる(日本テスト学会, 2010)。記述式の項目では，絶対的な正解がないような内容において，知識を応用する力をみたり，自らの考えをまとめ表現させたりすることができるというメリットがある一方で，

一度のテストの問題項目数が限られること，素材の選び方や公平性，出題のバランスに多枝選択式項目以上に気を付ける必要があること，採点作業に労力と時間がかかることがデメリットとして挙げられるという。選択式の項目では，一度に多くの項目を用意できること，出題内容のバランスがとりやすいこと，採点が容易であるなどというメリットがある一方で，得られる情報が選択枝に含まれる事項に限られること，選択枝自体が正答への手掛かりとなる受検者がいることが挙げられるという。

これらの項目形式の特徴を踏まえ，テストの目的，測定したい内容と照らしながら項目の形式を組み合わせることで，よいテストの作成へと繋がる。

テスト結果の利用・受検者へのフィードバックについて

作成したテストは実施され，そこで得られた解答は採点される。この採点の結果は，テストそのものや受検者に関しての多くの情報をもたらす。

テストの平均点や標準偏差，個々の受検者の得点をはじめ，そのテストの妥当性・信頼性に関する情報や，どの項目がうまく機能していたかなどの項目の質に関する情報まで入手することが可能である。そして，これらの情報を適切に扱うことで，次に作成・実施されるテストへと活かすことが求められる。

よいテストであったかどうかの判断をするためにも，これらの結果について検討を行うことが必要となる。すなわち，妥当性，信頼性といった側面から，設計通りのテストであったかどうかを確認することが求められる。加えて，合否に影響をするテストであれば，テストの質だけでなく，合否判断の質も求められる(光永, 2017)。つまり，その合否判断を行うカットオフポイントが妥当な判断であったのか等，合否という結果も含めてテストの評価がなされることが求められる。

また，クラスルームテストであれば，テストそのものの評価に加え，個々の受検者の正誤パタンといったテストから得られる情報を踏まえ，具体的に教育場面へ還元することも求められる。適切な分析を行えば，個々の受検者について，基礎が定着しているかどうかや，どの程度理解をしているかという情報を得ることが可能である(小泉, 2017)。そのためには，実施したテストの質が保証され，使用された項目について十分にその機能が確認されていることが必要となる。つまり，よいテストであることは，教育場面において有用な知見となりうる。

しかしながら，どれほど綿密に基本設計を練ったとしても，テスト項目それ自体を杜撰な方法で作成したり，テスト結果からフィードバックを得なければ，よいテストからは遠ざかる。つまり，基本設計を踏まえテスト項目を作成・評価が行われる必要がある。そこで次節以降では，よいテストを作成・評価するための具体的な手段として，項目作成ガイドラインならびに項目分析の手法について述べる。

2.2 よいテストを作るために―項目作成ガイドライン―

よいテストを作るためには，先述したようにテストの基本設計に則った形で作成する必要がある。しかしながら，どれほど入念に作成したとしても，テスト項目に存在する Flaw を完全に取り除くことは難しい。本節では, Flaw の影響を抑えるための方法の一つとして, 項目作成ガイドラインについて取り上げる。

項目作成ガイドライン

項目作成ガイドラインは，項目を作成する際のルールに関するリストである。

Haladyna & Rodriguez(2013)では，テスト項目の作成や，記述式テストの評価方法，アンケート調査の作成方法に至るまで, あらゆる形式に関する作成・評価のガイドラインを示している。特に多枝選択式項目作成ガイドラインは，Haladyna & Downing (1989)や Haladyna, Downing, & Rodriguez(2002)などで, 項目作成に関する検討を踏まえリスト化されている。また, こうして検討されてきたガイドラインは, 作成されたテストの質の確認に用いられたり (Jozefowicz, Koeppen, Case, Galbraith, Swanson, & Glew, 2002)，Flaw により学生の成績評価へ影響を及ぼすのかどうかの検討に用いられたりする(Tarrant & Hare, 2008)など，ガイドラインと Flaw に関する研究で多く用いられている。

多枝選択式項目作成についてのガイドラインでは，測定内容・配置・記述・設問部分・選択枝に関しての 27 のガイドラインが挙げられている。当該ガイドラインを日本語訳したものを表 2-1 に示す。テストの作問をするにあたり, これらのガイドラインを参照することで, Flaw の影響が抑えられた項目作成が可能となり，受検者の能力をより正確に測定できる。

しかし，同時に全てのガイドラインに則った項目を作成することは難しい。Breakall, Randles, & Tasker (2019) では，ガイドラインを作成し，それに基づいて既存の項目を改変したとしても，作成した全てのガイドラインに同時に則った項目は全体の 7.9%であり，多くの項目で準拠できないガイドラインが存在した。また, 問題の目的によっては，ガイドラインに則った項目を作成しないこともある。荒井 (2015) では，多枝選択式問題を作成している専門家へ作問時に気を付けていることの聞き取り調査を行い，その内容をガイドラインと照らし合わせている。その結果, ガイドラインでは用いない方がよいとされている項目形式について, 場合によってその形式を用いることが妥当であると考える作問者もいた。つまり，テストの目的や内容に応じてガイドラインに準拠するか否かについて柔軟な対応をする必要性が示唆されている。そのため, 問題の目的や項目の形式等に応じて，重視するべきガイドラインを選択する必要性がある。

表 2-1　多枝選択式項目作成ガイドライン(日本語訳)

測定内容に関して

1. 各設問は，ある1つの内容を理解し，記憶・解釈・応用等，ある1つの能力に基づいて解けること
2. より高次の能力を測る問題では，受検者にとって新奇な素材を用いること
3. 各設問の内容は互いに独立であること
4. 重要な事項を問うこと．極端に細かかったり，逆に一般的すぎる内容にならないこと
5. 解答が個人の意見に影響されないこと
6. ひっかけ問題にならないこと

配置に関して

7. とくに低学年の児童に対して，各選択枝は行を変えて1つずつ並べること

記述に関して

8. 内容面・形式面，また語法等について，よく校訂・校正すること
9. 言語レベルを受検者集団に合わせること
10. 設問・選択枝ともに，記述量を最小にすること．題意と無関係な文を入れないこと

設問部分に関して

11. その問題で何を問うているかは，設問部分に明確・簡潔に書くこと．選択枝を読まずとも，問題の意味が理解できること
12. 否定表現を使わないこと．もし使う場合は，否定表現部分を**強調表示**すること

選択枝に関して

13. もっともらしく，識別力の高い選択枝のみにすること．多くの場合，3枝で事足りる
14. 正答枝が唯一であること
15. 正答枝の位置をばらつかせること．系統性がないこと
16. 選択枝を，音順，数量の大きさなど，何らかの法則に従って配置すること
17. 各選択枝は互いに独立であること．内容に重なりがないこと．
18. 「上記のいずれでもない」「上記すべてあてはまる」「分からない」などの選択枝を用いないこと
19. 「でない」「～以外」などの否定表現を用いないこと
20. 正答枝を探す手掛かりを与えないこと
    a. 各選択枝の長さをおおむね揃えること
    b. 「絶対に」「常に」「決して」「完全に」など，強意語を用いないこと
    c. 設問と選択枝の間に，解答に影響するような語の重複や類似性がないこと
    d. 両立しない選択枝など，選択枝の内容から正答を絞り込めることのないようにすること
    e. 明らかに不要・不自然な選択枝は入れないこと
    f. 各選択枝の作りを等質にすること
21. どの誤答枝ももっともらしいこと．典型的な誤答を誤答枝に用いること
22. ユーモア（お遊び）は用いないこと

項目作成ガイドラインの影響

Downing (2005)は項目のガイドラインへの準拠／非準拠によって，合否への影響は少ないと指摘しているが，これは既に行われたテスト項目について検討されているものである。実施されたテストについて，ガイドライン準拠／非準拠項目に解答した受検者の正誤を用いて，ガイドラインの影響を検討している。つまり，ガイドライン準拠／非準拠項目は異なる項目について比較されており，項目間でガイドライン以外の条件が統制されていない。また，テストの目的は合否判断のためだけではない。

この他にも，ガイドラインについて検討したものとしては，Martínez, Moreno, Martín, & Trigo (2009) が挙げられ，ガイドラインに準拠しない項目では，ガイドラインに準拠している項目に比べ，正答率に差が生じるなどの影響があったとしている。しかし，これらの先行研究において，個々のガイドラインによる項目特性や受検者への影響についての検討はされているものの，複数のガイドラインを同時に網羅的に検討したものではない。つまり，目的に応じて重視するガイドラインを選択するためには，それぞれのガイドラインについて他のガイドラインとの比較により，正答率や識別力への影響力の大小を検討する必要がある。

ガイドラインへ影響するその他の要因

ガイドラインの影響力を検討するにあたり，ガイドラインに準拠していない項目について，受検者が違和感を覚えるかどうかを検討する必要もあると考えられる。上述したように，多くのテストにおいて Flaw のある項目が検出されているものの，これらは既に行われたテスト項目を用いた比較であり，それぞれの Flaw がどのような影響を及ぼすかどうかについて網羅的に行ったものではない。また，Flaw による項目特性への影響について検討されているものの，どのような Flaw が受検者に気付かれやすいのか，または気付かれにくいのかという検討は先行研究でなされていない。

項目の Flaw は測定道具としての機能を低下させる要因である。しかし，受検者にとって不利益となる Flaw があっても，その存在に受検者が気付くことができていれば，項目の不備を指摘する等，その場で対処することができると考えられる。一方で，受検者が Flaw に気付けない項目ではこのような対処は難しい。多くの受検者が Flaw に気が付かない項目でありながらも，ガイドラインに準拠しているか否かにより解答傾向が異なり，正答率が大きく変化する項目は，測定道具としての機能を十分に果たしていない。そのため，受検者が違和感を覚えないガイドラインについては，作題者は特に注意して作題する必要性があり，より重視するべきであると考えられる。

2.3 よいテストかどうかを評価するために―項目分析―

テストの解答が採点された後には，それがよいテストであったかを評価する段階へと進む。「能力の優劣の組織的な判定のための統計モデルの理論とその分析方法」をテスト理論という(大津, 2011)。ここではテスト理論について簡単に述べた後，よいテスト項目であったかを判断するための項目分析に焦点をあてテストの評価について述べていく。

テスト理論とその必要性

テスト理論とは，テストの標準化を目的としてテストやテスト得点を科学的対象として扱う分野である(荘島, 2010)。受検者の得点や正答率などを用いて検討をする古典的テスト理論と，項目に対する各受検者の正誤情報を数理モデル化し検討をする項目応答理論が主なものである。

古典的テスト理論(Classical Test Theory：CTT)は，線形モデルを用いてテスト項目の特性を分析するものである(大津, 2011)。CTT ではテストに関する統計量として，合計点の平均，標準偏差，相関係数などを用いる(池田, 1994)。CTT は，テストを構成する上で多くの技術的貢献を成した一方で，用いる統計量はテストの受検者集団の特性に影響を受けるという点に注意する必要がある。

項目応答理論(Item Response Theory：IRT)は，テスト項目への受検者の応答(e.g.正答―誤答)と，そのテストが測定しようとしている受検者の能力との関係に確率モデルを導入する理論である(野口, 2014)。IRT モデルのひとつに， 2 parameter logistic model(2PLM)がある。

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

これは，項目の識別力を表すa_jパラメタと，困難度を表すb_jパラメタという 2 つのパラメタによってモデル化されるものである。IRT を用いる利点としては，受検者の能力値と項目の困難度を同じ尺度上で検討可能であることや，得点と受検者の能力値を切り離して考えることができるという点である。また，同じ能力値をもつ受検者であるのに，所属している集団によってある項目の正答率が異なるといった特異項目機能(Differential Item Function :DIF)の影響を確認することもできる。一方，数理モデルとしてテストを扱う以上，そのモデルに適合するようなテストデータを用いることが必要となり，モデルから乖離する形式のテストや少人数のテストなどで利用することは難しい。

これらのテスト理論は，テスト場面に応じて使い分ける必要がある。幅広い母集団を想定した大人数に実施されるテストでは，IRT を用いることが望ましい場面もある一方で，受検者数が少なく，集団の特性が大きく変動しないクラスルームテストでは，CTT に基づいてテストの評価をすることの方が適していると考えられる。本論文では，CTT を理論的な下地として，テストの評価を行う方法について検討していく。

テストという測定道具の評価

テスト全体の評価，すなわちテストという測定道具についての評価は，まず信頼性・妥当性という側面で行うことができる。

信頼性の高いテストとは，テストから得られた解答に含まれている誤差が小さく，受検者の能力の真値をより反映したテストである。対して，信頼性の低いテストとは，テストから得られた解答に含まれている誤差が大きく，受検者の能力の真値をあまり反映しないテストである。

妥当性の高いテストとは，テストが測定したい概念を正しく測定できているテストのことである。対して，妥当性の低いテストとは，テストが測定したい概念をあまり測定できないテストのことである。主な妥当性の考え方としては，Messick(1989) や Kane(2013)が挙げられる。

これらの信頼性，妥当性による評価は，作成された「テスト」に対して行われる一方で，テストに含まれている個々の項目について検討することを，項目分析という。各項目の正答率，識別力を参照することでそのテスト項目が機能していたかどうかを検討し，以降のテスト項目作成へ有用な知見を与えることとなる。

項目分析の手法

項目分析では主に，正答率と識別力に焦点をあて検討していく。

テストの正答率とは，受検者の正答の割合を示すものである。これは，テストがどの程度難しかったかどうかを表す。項目の正答率は，回答者のうち，正答した受検者の割合で表すことができる。0 から 1 までの値をとり，0 のときすべての回答者が誤答したことを，1 のときすべての回答者が正答したことを表す。すなわち，正答率は 0 に近いほど難しく，1 に近いほど易しい項目であるという指標となる。

テストの識別力とは，そのテストがどの程度受検者の能力に基づいて得点を判別できるかどうかというものである。項目の識別力は，その項目がどの程度受検者の能力に基づいて正誤を判別できるかどうかというものである。そして，識別力指標とは，能力の低い受検者について正答率が低く，能力が高くなるほど正答率が上がる傾向を示す指標である(石井, 2007)。識別力は-1 から+1 の数値をとる。+1 のとき，識別力が最大であり，能力の高い受検者は正答するが，能力の低い受検者は誤答することを表す。0 は識別力がなしとして，受検者の能力によらず正答率が同じことを意味する。-1 では負の識別，すなわち，能力低群は正答するが能力高群では誤答することを意味する。

項目分析で用いられる識別力指標

識別力指標として，様々なものが検討されている。ここでは，I-T 相関，*D* 指標，IRT の

モデル上で扱われる識別力パラメタについて取り上げる。

I-T 相関は，ある項目における各受検者の正誤情報と，各受検者の合計得点の相関係数を算出することで識別力の指標とするものである。この指標では，項目数が少ないときには特に，合計得点から当該項目の正誤情報を引いた形で算出されることが望ましいとされる(野口, 2014)。なお，I-T 相関の算出については，石井(2020)などのソフトウェアを用いることにより，その値を得ることが可能である。しかし，I-T 相関を理解するためには点双列相関係数の理解が必要となり，かつ，相関関係を識別力として捉え直す必要がある。このように，I-T 相関そのものの概念的な理解が難しいことから，テストの非専門家が実施するテストにおいて扱うことは難しいとされる(Brown, 1996)。

D 指標(*D*-index)は，テストの受検者について能力値を上位群・中位群・下位群と分けた際の，上位群と下位群の正答率についての差をとったものである。上位群と下位群の差をとるという考え方は，Johnson(1951) で紹介されており，Ebel & Frisbie (1991) はこれが初出であるという。*D* 指標は簡便な識別力指標のひとつであるが，これは項目の正答率の影響を強く受ける(e.g.石井, 2007)。正答率が高い項目や低い項目では，十分に機能をしないという問題点がある。しかし *D* 指標は，上位群と下位群の正答率の差を取るのみで算出が可能であり，その数値も理解しやすい。

IRT の文脈では，識別力パラメタを識別力の指標として用いることが可能である。たとえば 2PLM では，項目の特性を表す 2 つのパラメタ(*a*パラメタ，*b*パラメタ)を想定している。項目間の a パラメタの値の大小関係を確認することで，項目特性曲線の変曲点付近における傾きを比較することができる。*a*パラメタの値が大きい項目において，受験者の能力値 θ による差異をより明らかに識別することができる。したがって，*a*パラメタは項目の識別力パラメタと呼ばれる(e.g.野口, 2014)。このパラメタは IRT モデルに基づく指標であるため，等化を行うことにより受検者に依存しない形での解釈が可能になるという利点がある。一方で，IRT の文脈で用いられるテストでしか算出することができないため，IRT の適用が難しいテストでは使用できず，場面が限られる。

クラスルームテストにおける項目分析

クラスルームテストでは，テスト実施後の評価としてテストの正答率(平均点)の算出は行われているものの，識別力についての検討はほとんど行われていない。しかし本来は，実施したテストが作成者の意に沿う出来であったかを検討するためにも，識別力を算出することが望ましい。

Diederich(1973)は，クラスルームテストでは非専門家がテストの評価を行うことを踏まえ，項目の識別力を評価する方法として簡便なアプローチが望ましいとし，その方法について提案している。これは，受検者をテストの合計得点から上位群と下位群の 2 群に分け，各項目の上位群の正答者数と下位群の正答者数の差を取ることで，項目間の識別力の大小を確

認するものである。また，Brown(1996)は言語テストにおける項目の識別力として，上位群と下位群の正答率の差を用いることを提案している。実際に，Wiyasa, Laksana, & Indrawati. (2019) や Laliamsyah. & Apriyanti (2020)などで，テストの項目分析としてこの方法に則り，項目識別力の算出が行われている。

先述した *D* 指標は簡便な識別力指標のひとつであるため，特にテストの非専門家が行うテストにおいては，有用な指標となり得ると考えられる。また，受検者能力群を横軸に，それぞれの群における正答率を縦軸にとり図示化する方法として，トレースラインがある。*D* 指標は，トレースラインのために受検者を三群に分けた際の上位群と下位群の高低差を数値化するものであり，視覚的情報との対応も容易である。このように，クラスルームテストにおける識別力指標として，*D* 指標を扱うことは簡便かつ有用であると考えられる。

D 指標を使用する際に生じる問題

D 指標を使用するにあたり，検討すべき点が 2 点存在する。1 点目は，受験者の群分けの割合について定められていないという点である。また，2 点目は，群分けのカットオフポイント上に複数の受検者が存在した場合の対応について，明確なルールが存在しないという点である。

まず，*D* 指標を用いるための群分けの割合に関する問題についてである。算出するためには受検者を 3 群に分ける必要があるが，それぞれの群の割合は定められていない。受検者の能力を分ける割合として，赤根・伊藤・林・椎名・大澤・柳井・田栗(2006)では，上位群/下位群を 25%ずつとした群分けを行っており，医学系試験の項目分析では多く用いられる割合だという。また，野口・大隅(2014)では，上位群・中位群・下位群の人数を揃えた例を紹介している。Kelly(1939) は，この割合について，上位群を 27%，下位群を 27%と提案している。これは，変数 X が正規分布するときの標本平均の差の臨界比を最大とする値に由来している。

次に，*D* 指標を用いる際に行う群分けの方法についても，十分な検討がなされていない。カットオフポイント上にいるテストで同点だった受検者について，同じ群とするか異なる群とするかによっても *D* 指標は異なる値となる。また，同点だった受検者について，Kelly(1939)のような割合になるように異なる群に割り振るのであれば，受検者を分ける方法は多様に考えられる。このとき，テスト全体の *D* 指標は群分けの方法によらず同じ値を示すが，テストに含まれる項目の *D* 指標は群分けの影響を受けることになる。適切な群分けが行われていなければ，そこで表される *D* 指標は群の分け方という誤差を含んだものとなる。同点の受検者を減らすために，テストの項目数を増やすことも考えられるが，学校現場などで実際に用いられるテストではテスト時間に制限があったり，作問する教員への負担になったりするなど限界がある。そのため，識別力の高い項目のみで構成されたテストであっても，同点の受検者をゼロにすることは容易ではない。

これらのことを踏まえ，クラスルームテストにおいて D 指標を扱うためにも，どのように 3 群を分けることでより受検者の能力を反映した形で D 指標を算出することが可能かについて検討をする必要がある。このとき，より簡便な基準を用いて群分けを行うことが望ましいと考えられる。

第3章

テストの作成に関する研究[2]

――多枝選択式項目作成ガイドラインに関する実証的研究（研究 1）

[2] 本研究は，坪田・石井(2020)を改稿したものである。

3.1 問題と目的

テスト項目は，測定したい能力に応じてその正誤が判別されることが求められるが，項目の Flaw は受検者の能力以外によって正誤に影響を及ぼす要因になりうる。こうした Flaw の影響を低減させるために，項目作成ガイドラインが提案されている。項目作成ガイドラインを用いることで，項目の Flaw は抑えられるものの，どの程度の影響があるのかについては実証的に検討されていない。また，こうした Flaw が存在する場面では，受検者がどのように対処するべきかについては検討されておらず，そもそも受検者が Flaw に気が付くかについても検討する必要がある。

項目形式のひとつである多枝選択式テストは，多くのテスト場面で用いられており，あらかじめ複数用意された選択枝から適切なものを選ぶという形式 (日本テスト学会, 2007) のものである。大学入試センター試験や大学入学共通テストなどの大規模試験だけでなく，個別学力試験等でも用いられることは多い。実際に，国立大学の一般入学個別学力試験でも総項目数の 12.5%は客観的項目であり，かつそのうちの 75%は多枝選択式形式を用いていた (宮本・倉元, 2017)。このように，多枝選択式形式は試験規模を問わず多く用いられている。

ところで，Tarrant & Ware(2008)では，Flaw のある項目について，難易度が高くなるものや低くなるものなど，一貫性のない結果を示したといい，Flaw が回答に様々な影響を及ぼしているという。正答がない，もしくは複数あるという Flaw が確認された項目では，Flaw のない項目たちに比べ難易度が上がり，反対に，正答への手掛かりに関する記述が多いという Flaw をもつ項目では，難易度は下がったという。このように，一言で Flaw のある項目といっても，その影響は一貫していない。また，Flaw のある項目と Flaw のない項目を比較している文献においても，同じ能力を測定している項目について検討されたものではない。そのため，Flaw が及ぼす量的な影響について言及することはできない。

研究 1 では，多枝選択式項目について，項目作成ガイドライン(Haladyna & Rodriguez, 2013)の影響について検討することを目的とする。同一の能力を測定する項目として項目作成ガイドラインに準拠する作問をした項目と準拠しない作問をした項目の正答率や識別力について網羅的に比較検討する。それにより，それぞれのガイドラインに準拠した際にどの程度影響を与えるのかを明らかにする。加えて，受検者がガイドラインや Flaw にどの程度気が付くのかを検討することで，項目へ影響を与えながらも，受検者に気付かれない(意識することが難しい)ガイドラインを明らかにする。

3.2 テストの作成と実施

テストの作成

テスト項目として，Haladyna, & Rodriguez (2013) の多枝選択式項目作成ガイドライン(表2-1)に基づき，52 項目を新規作成した。27 のガイドラインのうち，25 のガイドラインについて，準拠項目 1 項目と非準拠項目 1 項目の 2 項目を作成した。ただし，ガイドライン 3 は項目間の独立性に関するガイドラインであったため，このガイドラインのみ準拠項目 2 項目と非準拠項目 2 項目の計 4 項目を作成した。作成した項目について，教育測定学を専門とする大学院生と教員の 2 名で検討をした後，心理学を専攻する 3 名の研究協力者に予備調査を行った。予備調査で得られた意見を踏まえ，再度検討を行い，実施項目ならびに実施時間を確定した。

なお，項目作成から除外した 2 つのガイドラインについて，ガイドライン 7「とくに低学年の児童に対して，各選択枝は行を変えて 1 つずつ並べること」では，調査対象者が大学生であったため除外した。また，ガイドライン 15「正答枝の位置をばらつかせること．系統性がないこと」はテスト全体の正答枝の作りについてのガイドラインであるため，本研究では検討対象としなかった。

本調査では, 実験用冊子 2 種類とアンカーテスト 1 種類の計 3 種類のテストを使用した。それぞれのテスト冊子には全て多枝選択式問題で，国語 8 項目，数学 6 項目，英語 12 項目の全 26 項目を収録した。ガイドライン項目と教科の対応は，表 3-1 である。

表 3-1 教科とガイドライン項目の対応

教科	ガイドライン番号
国語	2,13,19,20a,20b,20c,20d,20e
数学	1,3,9,10,11,12
英語	4,5,6,8,14,16,17,18,20f,21,22

実験用冊子では，ガイドラインに準拠する項目 (以下，準拠項目) ，準拠しない項目 (以下，非準拠項目) を，13 項目ずつランダムに割りあてた。アンカーテストは 26 項目全てに準拠項目を用いた。

受検者

2019 年 5 月に愛知県内の国立および私立大学生 477 名にテストを行った。4 つの大学でテストを実施し，受検者はトップ～中堅レベルの大学 1～4 年生，大学院生であった。母語が日本語ではない者，および全ての項目に同じ解答をしている者を除いた有効回答者数は 453 名であった。

手続き

テストは授業時間後に自由参加として行われた。テストに参加するにあたり，報酬等はなかった。

受検者には 3 種類のテスト冊子のうち 1 種類をランダムに配布した。テスト冊子が全員に配布されたことを確認した後，表紙に記載している事項を口頭で説明した。その上で，それぞれの項目の正答枝を選ぶよう求め，一斉にテストを開始した。テスト実施時間内にテスト全体や項目等について気が付いた点について，自由記述で回答するように求めた。質問は一切受け付けず，何かあれば自由記述欄に記述をするよう教示を行った。

なお，問題に対する解答ならびに自由記述の回答はすべて解答用紙に記入するよう求めた。

テスト実施時間は 30 分間とした。

無回答の扱い

数学にのみ解答しなかった 10 名の受検者について，数学の項目のみを無回答として扱った。それ以外の無回答は誤答として扱った。

3.3 全体的な結果

本研究では，受検者が解答した冊子によって解答項目が異なっていた。

準拠項目である 26 項目について，分散・共分散行列を用いて α 係数を推定したところ，α=.79 であった。非準拠項目は共通受検者が存在しないため，共分散行列を推定することができない。しかし，同じガイドラインについての準拠項目と非準拠項目で同一の概念を測定している。非準拠項目を含んでいる冊子 1，冊子 2 の α 係数はそれぞれ，α=.70, .74 であった。そのため，全体として十分な信頼性と判断した。

異なる冊子に解答した受検者の能力値について比較を行うため，準拠項目と非準拠項目のそれぞれについて，解答しなかった項目を無回答とし，52 項目のテストと見なした。ガイドライン準拠項目と非準拠項目を同時に検討するため，1PLM を用いて受検者の能力パラメタ θ を算出した。各項目の正誤と受検者の能力推定値 θ の相関 (Item-Theta 相関) を項目識別力の指標として用いた。

各ガイドライン項目の I-T 相関，選択枝選択率，コメント率を表 3-2 に示した。

コメント率

解答用紙裏面の自由記述のうち，テスト項目に対して言及されたものをコメントとして扱った。それぞれの項目へのコメントは，ガイドラインに関するものか否かを問わず，全てのコメントを抽出した。

受検者の中で 1 つ以上のコメントをした者は 242 名で全体の 53.4%であった。受検者の能力値 θ とコメント数についての相関係数は，r=.27 であり，能力値とコメント数の間に関連は見られなかった。θ とコメント数の散布図を図 3-1 に示した。

正答率

ガイドライン準拠／非準拠により，正答率が大きく変化したガイドラインの存在が確認された。準拠項目全体の正答率の平均は 0.61 であり，SD は 0.21 であった。一方，非準拠項目全体の正答率の平均は 0.56 であり，SD は 0.24 であった。同一ガイドラインを参照して作成した準拠／非準拠項目について，正答率の差を算出したところ，平均は 0.05 であり，SD は 0.17 であった。

識別力

ガイドライン準拠／非準拠により，I-T 相関の値が大きく変化したガイドラインの存在が確認された。準拠項目全体の I-T 相関の平均は 0.36 であり，SD は 0.12 であった。一方，非準拠項目全体の I-T 相関の平均は 0.34 であり，SD は 0.16 であった。同一ガイドラインを参照して作成した準拠／非準拠項目について，I-T 相関の差を算出したところ，平均は 0.01 であり，SD は 0.11 であった。

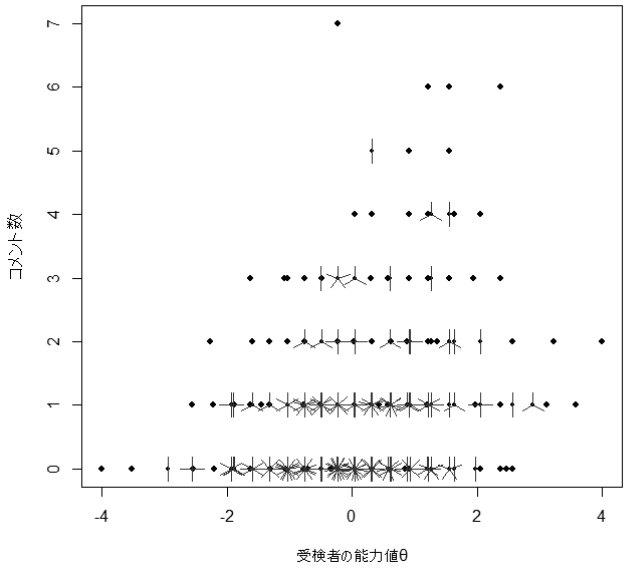

図 3-1　受検者の能力値とコメント数の散布図

検討除外ガイドライン

ガイドライン 20b について，ガイドライン非準拠項目に強意語として不十分な選択枝が存在したため，検討から除外した。また，ガイドライン 22 について，ガイドライン準拠項目に正答選択枝と見なせる項目が複数存在したため，検討から除外した。

表 3-2　各ガイドライン項目の選択枝選択率，I-T 相関，コメント率

		選択枝1	選択枝2	選択枝3	選択枝4	I-T相関	コメント率
1. 各設問は，ある1つの内容を理解し，記憶・解釈・応用等，ある1つの能力に基づいて解けること		0.17	0.33	0.33	0.11	0.13	0.0%
		0.22	0.35	0.21	0.16	0.15	4.3%
2. より高次の能力を測る問題では，受検者にとって新奇な素材を用いること		0.05	0.13	0.71	0.10	0.46	0.4%
		0.05	0.14	0.70	0.11	0.39	0.6%
3. 各設問の内容は互いに独立であること	1	0.08	0.13	0.66	0.09	0.46	0.7%
	2	0.08	0.48	0.29	0.11	0.49	0.0%
	1	0.08	0.10	0.73	0.06	0.41	0.0%
	2	0.39	0.13	0.20	0.23	0.59	0.6%
4. 重要な事項を問うこと．極端に細かかったり，逆に一般的すぎる内容にならないこと		0.16	0.21	0.29	0.33	0.14	1.9%
		0.13	0.22	0.30	0.34	0.28	26.1%
5. 解答が個人の意見に影響されないこと		0.03	0.86	0.05	0.05	0.09	14.2%
		0.65	0.10	0.17	0.07	-0.28	35.4%
6. ひっかけ問題にならないこと		0.01	0.05	0.25	0.68	0.23	1.8%
		0.02	0.08	0.22	0.66	0.35	7.9%
8. 内容面・形式面，また語法等について，よく校訂・校正すること		0.13	0.10	0.54	0.23	0.45	0.0%
		0.19	0.06	0.50	0.25	0.42	0.0%
9. 言語レベルを受検者集団に合わせること		0.09	0.54	0.19	0.16	0.42	0.7%
		0.14	0.52	0.13	0.15	0.34	62.9%
10. 設問・選択枝ともに，記述量を最小にすること．題意と無関係な文を入れないこと		0.49	0.11	0.27	0.12	0.46	0.0%
		0.52	0.11	0.21	0.12	0.42	12.0%
11. その問題で何を問うているかは，設問部分に明確・簡潔に書くこと．選択枝を読まずとも，問題の意味が理解できること		0.22	0.50	0.21	0.06	0.43	0.0%
		0.18	0.56	0.18	0.06	0.39	4.9%
12. 否定表現を使わないこと．もし使う場合は，否定表現部分を**強調表示**すること		0.08	0.12	0.55	0.25	0.23	0.0%
		0.06	0.13	0.59	0.23	0.26	0.5%
13. もっともらしく，識別力の高い選択枝のみにすること．多くの場合，3枝で事足りる		0.20	0.72	0.08		0.38	1.5%
		0.13	0.70	0.13	0.04	0.50	0.0%
14. 正答枝が唯一であること		0.09	0.05	0.04	0.81	0.40	0.4%
		0.09	0.04	0.03	0.82	0.34	1.1%
16. 選択枝を，音順，数量の大きさなど，何らかの法則に従って配置すること		0.03	0.03	0.91	0.03	0.41	0.0%
		0.04	0.91	0.03	0.02	0.43	0.0%
17. 各選択枝は互いに独立であること．内容に重なりがないこと．		0.01	0.97	0.01	0.00	0.36	0.7%
		0.01	0.90	0.01	0.06	0.23	17.4%
18. 「上記のいずれでもない」「上記すべてあてはまる」「分からない」などの選択枝を用いないこと		0.34	0.10	0.22	0.34	0.35	1.5%
		0.08	0.49	0.04	0.39	0.27	3.3%
19. 「でない」「〜以外」などの否定表現を用いないこと		0.60	0.15	0.15	0.10	0.50	0.7%
		0.35	0.37	0.20	0.08	0.48	2.2%
20. 正答枝を探す手掛かりを与えないこと							
a. 各選択枝の長さをおおむね揃えること		0.07	0.64	0.24	0.05	0.49	0.0%
		0.06	0.65	0.15	0.14	0.46	2.2%
b. 「絶対に」「常に」「決して」「完全に」など，強意語を用いないこと		0.05	0.22	0.11	0.61	0.44	0.4%
		0.08	0.51	0.13	0.27	0.44	2.7%
c. 設問と選択枝の間に，解答に影響するような語の重複や類似性がないこと		0.16	0.60	0.10	0.14	0.29	1.5%
		0.16	0.61	0.07	0.15	0.25	0.0%
d. 両立しない選択枝など，選択枝の内容から正答を絞り込めることのないようにすること		0.05	0.16	0.54	0.23	0.19	0.7%
		0.05	0.05	0.61	0.27	0.32	0.5%
e. 明らかに不要・不自然な選択枝は入れないこと		0.09	0.16	0.52	0.22	0.48	0.0%
		0.12	0.24	0.58	0.05	0.45	3.9%
f. 各選択枝の作りを等質にすること		0.03	0.01	0.91	0.05	0.46	0.0%
		0.04	0.02	0.93	0.01	0.34	0.6%
21. どの誤答枝ももっともらしいこと．典型的な誤答を誤答枝に用いること		0.59	0.12	0.14	0.15	0.28	0.4%
		0.66	0.09	0.12	0.12	0.35	0.5%
22. ユーモア（お遊び）は用いないこと		0.09	0.05	0.04	0.81	0.40	0.4%
		0.09	0.04	0.03	0.82	0.34	1.1%

網掛けが非準拠項目
下線が正答選択枝

3.4 各ガイドラインの検討

ここでは，ガイドラインに準拠しているか否かにより I-T 相関，正答率，コメント率のいずれかへの影響が大きく，特に注意するべきガイドラインについて検討する。

ガイドライン 3「各設問の内容は互いに独立であること」

本ガイドラインを検討する項目は，小問 2 題((1),(2)と以下表記)で構成された。(1)では，二次関数の式を求めさせ，(2)では二次関数より，指定された範囲の最小値を求めさせた。準拠項目では，これら 2 題を独立な項目として出題した。一方，非準拠項目では，(1)で求めた二次関数を用いて，(2)で最小値を求める必要があった。使用した項目を表 3-3 に示す。

表 3-3　ガイドライン 3 として用いた項目

3. 各設問の内容は互いに独立であること	
準拠	(1). 3点(0,3)，(-2,17)，(1,5)を通る二次関数$f(x)$を求めよ。 1.　$f(x)=x^2+x+3$ 2.　$f(x)=2x^2+3$ 3.*　$f(x)=3x^2-x+3$ 4.　$f(x)=4x^2-2x+3$ (2). 二次関数$y=2x^2-5x+5$の範囲が$-3 \leqq x \leqq 4$のとき，yの最大値と最小値の組み合わせとして正しいものを選べ 1.　35/16 2.*　15/8 3.　17 4.　38
非準拠	(1). 3点(0,3)，(-2,17)，(1,5)を通る二次関数$f(x)$を求めよ。 1.　$f(x)=x^2+x+3$ 2.　$f(x)=2x^2+3$ 3.*　$f(x)=3x^2-x+3$ 4.　$f(x)=4x^2-2x+3$ (2). (1)で求めた二次関数の範囲が$-3 \leqq x \leqq 4$のとき，yの最小値として正しいものを選べ 1.*　35/12 2.　28/9 3.　47 4.　33

(1)の正答率では非準拠項目が 0.73 であり，準拠項目の 0.66 よりも高かったのに対し，(2)の正答率は 0.39 と準拠項目の 0.48 よりも低かった。また，(1)の影響を受ける(2)の I-T 相関は非準拠項目で 0.59 であり，準拠項目の 0.49 よりも高かった。項目に依存性がある場合，項目識別力が上昇することは先行研究でも指摘されており (Chen & Wang, 2007)，本研究でも同様の結果となった。(1)での誤答が(2)の正答を許さないため，(1)を正答した一定以上の能力値の受検者のみ(2)に正答でき，結果的に識別力が高くなったと考えられる。つまり，項目間に依存性が生じているとき，識別力が不当に高まる。

項目に対してのコメント率は，準拠／非準拠にかかわらず低い。また，非準拠項目のコメントの内容も，「よくわからなかった」というものであり，問題形式に対してのコメントではなかった。このことから，受検者はこうした問題形式に違和感を覚えないことが明らかとなった。

本ガイドラインに準拠しないことで，項目の形式という本来測りたい特性以外の要因により，識別力は高くなり，正答率が低下する。その上，受検者はその形式が正答率を低下させる要因となっていることに気が付いていない。そのため，本ガイドラインは受験生の正答率へ及ぼす影響が大きく，重視するべきガイドラインであると考えられる。

ガイドライン 4「重要な事項を問うこと。極端に細かかったり，逆に一般的すぎる内容にならないこと」

本ガイドラインを検討する項目は，英単語について意味を選択する項目であった。準拠項目では，大学受験レベルの英単語を出題し，非準拠項目では専門性が高く極端に難易度の高い英単語を出題した。使用した項目を表 3-4 に示す。[3]

準拠項目の正答率は 0.21 であり，非準拠項目の正答率は 0.22 と，正答率の差はほとんど見られなかった。しかし，非準拠項目ではコメント率が 26.1%と高く，「見たことのない単語」や，選択枝の内容から「こんな単語を知っている必要性を感じられない」という趣旨のコメントが多い一方で，一部の受検者は，「英単語の語源を含め検討すると正答選択枝を絞ることができた」というコメントを残していた。

本項目では，非準拠項目において準拠項目よりも I-T 相関が高かった (非準拠 0.28，準拠 0.14)。この理由として，能力値の高い受検者にとっては今回出題した単語の中に，正答選択枝を導くヒントを見つけられたが，準拠項目では単純に単語の意味を知っているか否かが

[3] なお，本ガイドラインにて用いた準拠項目の選択枝 3「古門書学」は正しくは「古文書学」と誤植があったが，他の誤答枝と比較して選択率に大差がなく，また本ガイドラインについて検討する上で影響のない誤りであることから，項目分析への影響は少ないと判断した。

正誤に影響した可能性が示唆された。本項目の目的が「英単語の意味を知っていること」である以上，非準拠項目での解答傾向は好ましくないものである。そのため，本ガイドラインは受検者に違和感を覚えさせる上に，測定したい能力とは別次元の能力を測定する可能性もあることから，準拠したいガイドラインであると考えられる。

表 3-4　ガイドライン 4 で用いた項目

4. 重要な事項を問うこと．極端に細かかったり，逆に一般的すぎる内容にならないこと	
準拠	以下の単語の意味を選びなさい。 **diploma** 1.　交渉術 2.* 卒業証書 3.　古門書学 4.　外交官
非準拠	以下の単語の意味を選びなさい。 **pneumonoultramicroscopicsilicovolcanoconiosis** 1.　肺炎 2.* 塵肺症 3.　一過性脳虚血発作 4.　副甲状腺機能亢進症

ガイドライン 5「解答が個人の意見に影響されないこと」

本ガイドラインを検討する項目は，英文を読み，その文章が示す適切な人物の選択枝を解答する項目であった。準拠項目では，解答が客観的に決まっている内容であったが，非準拠項目では解答が主観的に決められていた。使用した項目を表 3-5 に示す。

非準拠項目において，受検者は解答を導き出す術を持ち得ておらず，I-T 相関が-0.28 と負の値になった。多くの受検者は「人によって正答が異なる」という旨のコメントをしていた一方，「英文で書かれているテーマについての知識が不足しているため正答選択枝を選ぶことができない」というコメントも少なからずあった。受検者は前提としてテスト項目に Flaw があることを疑わない人もいると考えられる。

また，非準拠項目において最も選択率の高かった選択枝 1 の I-T 相関は 0.12 であった。つまり，最も選択率の高い項目を正答枝としたとしても，受検者の能力と本項目の正誤には関連がない。

このことから，作題者の主観で正答枝を決めたり，そうした項目で選択率が最も高い選択枝を正答枝としたりすることは受検者の能力を反映しない項目となることが示唆された。

表 3-5　ガイドライン 5 で用いた項目

5. 解答が個人の意見に影響されないこと	
準拠	( )に当てはまるものを選べ。 ( ) is the host of Tokyo Disney Resort. 1. Leonardo DiCaprio 2.* Mickey Mouse 3. Harrison Ford 4. Simon Baker
非準拠	( )に当てはまるものを選べ。 ( ) is the best actor in the world. 1. Leonardo DiCaprio 2.* Mickey Mouse 3. Harrison Ford 4. Simon Baker

ガイドライン 9「言語レベルを受検者集団に合わせること」

本ガイドラインを検討する項目は，中学生相当の難易度の数学の項目であった。準拠項目では，言語レベルを受検者に合わせた記述をしたが，非準拠項目では同様の内容の設問について言語レベルを下げ，全て平仮名で記述して出題した。使用した項目を表 3-6 に示す。

表 3-6　ガイドライン 9 で用いた項目

9. 言語レベルを受検者集団に合わせること	
準拠	半径7の円Pと半径5の円Qが異なる2点で交わり，二つの円の中心間の距離をdとする。このとき，dの取りうる値の範囲を求めよ。 1. $d < 2, 12 < d$ 2.* $2 < d < 12$ 3. $d \leqq 2, 12 \leqq d$ 4. $2 \leqq d \leqq 12$
非準拠	はんけい 7の　えん い と　はんけい 5の　えん ろ が　ことなる　にてんで　まじわり　ふたつの　えんの　ちゅうしんかんのきょりを は とする このとき　は の　とりうるあたいの　はんいを　もとめよ。 あ.　は しょうなり 2　　12 しょうなり は い.*　2 しょうなり は しょうなり 12 う.　は しょうなりいこおる 2　　12 しょうなりいこおる は え.　2 しょうなりいこおる は しょうなりいこおる 12

非準拠項目でコメント率が 62.2%と非常に高く，受検者の多くが全て平仮名の記述に対し疑問を呈していた。しかしながら，「全て平仮名を用いることで識字障害等への配慮が見られた」等の肯定的な意見を記述した受検者も数名いた。適切な記述をした準拠項目と比べて識別力は若干低下したものの，正答率はほぼ同じであった。そのため，多くの受検者は誠実に対応したと考えられる。

一方で，本項目で非準拠項目に解答した者のみ，次の設問の正答率が大幅に低くなった。漢字仮名交じりの記述と全て平仮名での記述では，読みやすさに及ぼす文脈的・意味的要因が異なる (北尾, 1960)。全て平仮名で記述したとき，誤読数が多く，繰り返し読むことで誤読数が減る。また，漢字は視覚入力されても音韻変換することなく意味理解をするのに対して，平仮名では漢字よりも深い処理が必要になるとされている (篠塚・窪田, 2012)。そのため，本項目の解答にあたり，他の設問に比べて時間ならびに認知資源を多く費やし，次の設問に十分に解答できなかった可能性が考えられる。

本ガイドラインの検討項目では，非準拠項目において受検者に対して大幅に言語レベルを下げた。そのため，多くの受検者は設問を理解することは可能であったと考えられる。このことが, 正答率への影響が小さかった (準拠 0.54, 非準拠 0.52) 一因であると考えられる。一方，受検者に対して言語レベルを上げたとき，問題文や選択枝の理解が不十分となり，正答率や識別力に影響が出る可能性がある。

つまり，本ガイドラインは言語レベルを受検者より上げるだけでなく下げたとしても，後の項目に影響を及ぼすことが明らかとなった。

ガイドライン 17「各選択枝は互いに独立であること。内容に重なりがないこと」

本ガイドラインを検討した項目は，問題文の英文が説明する単語として適切なものを選択する項目であった。準拠項目では，異なる 4 つの選択枝が用意されていたのに対し，非準拠項目では，正答選択枝と包含関係になる選択枝が一つ存在していた。具体的には，正答枝は一般名詞であり，もう一つの選択枝は固有名詞であった。使用した項目を表 3-7 に示す。

非準拠項目において準拠項目よりも正答率は低く （準拠項目 0.97，非準拠項目 0.90），独立でないもう一つの選択枝 4 の選択率が独立である準拠項目の選択枝 4 よりも高かった (準拠 0.00，非準拠 0.06)。当該項目は準拠／非準拠項目のどちらも困難度が低く，多くの受検者が正答選択枝を絞ることが可能であった。しかし，準拠項目では唯一に絞ることができる正答枝に対し，非準拠項目では 2 枝から絞ることができず，どちらも解答欄に書いた受検者が存在した。非準拠項目ではコメント率も 17.6%と高く，内容も，「選択枝 2 と 4 はどちらも正しい」というものが多かった。解答を唯一に絞る過程では，「どちらも正答である」から両方の選択枝を記述した受検者と，「どちらも正答であるが，包含関係であるより大きなカテゴリ」である正答枝のみを記述した受検者に分かれた。池上(2015) は，言語テストで

観察されたテストテイキングストラテジーとして，採点基準等を想像して丸をもらえそうな選択枝を探すという出題者の意図を推測するものがあったとしている。今回，どちらも正答である選択枝の中で 1 枝に絞ることができた者とできなかった者の間には，このようなストラテジーを持っていたか否かが影響している可能性がある。

表 3-7　ガイドライン 17 で用いた項目

17. 各選択枝は互いに独立であること．内容に重なりがないこと	
準拠	以下の説明が示すものを選びなさい。 a shop where you can buy food, alcohol, magazines etc, that is often open 24 hours each day 1. Park 2.* Convenience store 3. City hall 4. Fire station
非準拠	以下の説明が示すものを選びなさい。 a shop where you can buy food, alcohol, magazines etc, that is often open 24 hours each day 1. Park 2. * Convenience store 3. City hall 4. Lawson

表 3-8　ガイドライン 14 で用いた項目

14.正答枝が唯一であること	
準拠	（ ）に当てはまるものを選べ。 Alan Menken ( ) is a composer won an Academy Award for the Original Music Score of “Beauty and the Beast”. 1. whom 2. which 3. what 4.* who
非準拠	（ ）に当てはまるものを選べ。 Alan Menken ( ) is a composer won an Academy Award for the Original Music Score of “Beauty and the Beast”. 1.* that 2. which 3. what 4.* who

一方，ガイドライン 14「正答枝が唯一であること」の非準拠項目にも正答選択枝が 2 枝あるが (表 3-8)，コメント率は 1.1%と多くの受検者は気が付かなかった。もう一つの正答選択枝である選択枝 1 の選択率も 0.09 と低く，準拠項目の誤答選択枝 1 とほぼ同じ選択率であった。また，選択枝 1 と 4 を同時に解答した受検者も存在しなかった。

ガイドライン 14，17 の非準拠項目では，どちらも正答選択枝が 2 つ存在している。しかし，コメント率は大きく異なっていた。この理由として，ガイドライン 14 の正答選択枝は受検者にとって正答であるということがわかりやすく，一方でガイドライン 17 の正答選択枝は唯一の正解であるということがわかりにくかったことが考えられる。選択枝に 2 つ正答枝があったとしても，どちらを正解と出題者が考えているかがわかりやすいとき，受検者は正答枝を唯一に決めて解答をする。一方で，どちらの選択枝が正答となっているかわかりにくいとき，受検者は 2 つの選択枝で迷うことから，項目の不備が認知されることで，コメントすると考えられる。したがって，ガイドライン 17 では選択枝 2 と 4 のどちらを正答と見なされるかが曖昧であったため，コメント率が高くなったと考えられる。

正答選択枝と捉えることができる選択枝が複数存在しているとき，テストテイキングストラテジーを用いることの可否により正答となるかどうかに影響する可能性があり，どの選択枝を正答選択枝とするかにより正答率や識別力に影響を及ぼす。そのため，正答選択枝になり得る選択枝が複数存在する項目は避けたほうが良いことが実証された。

ガイドライン 18「『上記のいずれでもない』『上記すべてあてはまる』『わからない』などの選択枝を用いないこと」

本ガイドラインを検討する項目は，複数の意味をもつ英単語の意味を知っているかを問うものであった。準拠項目では，提示した語には含まれない意味を選択するものであり，非準拠項目では，当てはまる意味を答えさせる項目とし，選択枝 4 として「上記すべてあてはまる」を用いた。使用した項目を表 3-9 に示す。

「上記すべて当てはまる」という選択枝は，全ての選択枝について正誤を検討する必要があるように見える形式であるが，準拠項目よりも非準拠項目において I-T 相関は低く (準拠 0.35，非準拠 0.27)，正答率は高かった (準拠 0.34，非準拠 0.39)。

非準拠項目の解答傾向に着目すると，選択枝 2 のみが正しいと考える受検者が多く，その他，選択枝 1 か 3 のどちらかについても正しいとわかった受検者は選択枝 4 を選んでいると考えられる。一方，選択枝 1 や 3 のみが正しいと考える受検者は少なく，それぞれ低い選択率となっている。

準拠項目では，同一英単語について，その単語が持たない意味を選択させた。選択枝 1〜3 は非準拠項目と同一の意味を用い，選択枝 4 を誤答選択枝とした。解答傾向に着目すると，選択枝 2 が当該英単語のもつ意味だと知っている受検者は多く，選択枝 2 の選択率は 0.10 と低い。選択枝 1，3 についても正しいとわかった受検者は選択枝 4 の正答枝を選ぶ一方，

選択枝 1 か 3 のどちらかについてのみ正しいとわかった受検者はそれ以外の選択枝と選択枝 4 のどちらが正答選択枝かわからなかったと考えられる。そのため，選択枝 1 の選択率が 0.34，選択枝 3 の選択率が 0.22 と高くなり，非準拠項目に比べて正答率が低下したと考えられる。

表 3-9　ガイドライン 18 で用いた項目

18.「上記のいずれでもない」「上記すべてあてはまる」「分からない」などの選択枝を用いないこと	
準拠	“issue”の意味として正しくないものを選べ。 1.　発行する 2.　問題 3.　(雑誌などの)号 4.*　学説
非準拠	“issue”の意味として正しいものを選べ。 1.　発行する 2.　問題 3.　(雑誌などの)号 4.*　上記すべてあてはまる

準拠項目，非準拠項目のどちらでも，選択枝 2 は提示された単語のもつ意味であると理解している受検者が多く，選択枝 1，3 の意味まで知っている受検者はそれほど多くなかったと考えられる。つまり，受検者の持っている知識体系はほぼ同じであるにも関わらず，準拠項目と非準拠項目で特性値が変化した。

4 枝選択式において，1 選択枝を「上記すべてあてはまる」とすることは，残りの 3 枝のうち 2 枝が当てはまるとわかった時点で，正答を絞ることが可能である。同様のことが「上記のいずれでもない」「わからない」にも当てはまる。つまり，全ての選択枝について精査することなく正答選択枝を導き出すことが可能になる。

全ての選択枝について検討させるために，選択枝数を増やしたり，複雑な組み合わせ形式を用いたりすることが考えられる。実際に医師国家試験では，複数の正答枝を選択させる形式の項目が出題されている。しかし，正答選択枝を全て選ばせたとき，学生の多くは確実に正答と理解する 1 選択枝を選ぶに留まる傾向がある (遠山・中村, 2013) ことや，テストワイズネスを助長させるだけという批判 (斎藤・有田, 1981) もある。また，同一内容を問う複数のテストにおいて，問題形式によって正答に必要な受検者の知識量レベルはほぼ同じ分布をするのに対し，複数選択式では識別力が高くなる項目が多い (木村・福島・栗原・黒澤,

2000)。つまり，解答するために必要な知識量にはあまり差はなく，いたずらに識別力を上げるに留まり，より高次な能力を測るものではない。高すぎる識別力は，一部の上位層のみを選抜することになり，大規模テストのような幅広い受検者層を想定したテストでは好ましくない。

したがって，多枝選択式を用いるにあたっては，ガイドラインでも指摘されているように，唯一である正答選択枝を選択させることが項目として最大限に機能するといえる。

ガイドライン 19「『でない』『～以外』などの否定表現を用いないこと」

本ガイドラインを検討する項目は，和歌で用いる修辞法について問う項目であった。準拠項目では，選択枝に否定表現は用いず，間違っている選択枝を選ぶ項目であった。非準拠項目では，選択枝の文言の中に多くの否定表現が含まれており，正しいものを選択する項目であった。実際の項目を表 3-10 に示す。

表 3-10　ガイドライン 19 で用いた項目

19.「でない」「～以外」などの否定表現を用いないこと	
準拠	和歌の修辞法について**誤っているもの**を選べ。 1.* 句切れとは，第五句の終わりに意味上の切れ目があるものである。 2. 枕詞とは，主に5音で特定の語を導き出すことを目的とした決まった言葉である。 3. 縁語とは，ある言葉と意味の上で関連する言葉を連想的に用いる技法のことである。 4. 本歌取りとは，有名な古歌から1～2句を取って新しい歌を詠む技法である。
非準拠	和歌の修辞法について正しいものを選べ。 1.* 句切れとは，第五句以外の句の終わりに意味上の切れ目があるものである。 2. 枕詞とは，主に5音以外で特定の語を導き出すことを目的とした決まった言葉である。 3. 縁語とは，ある言葉と意味の上で関連しない言葉を連想的に用いる技法のことである。 4. 本歌取りとは，有名でない古歌から1～2句を取って新しい歌を詠む技法である。

項目の内容は準拠項目と非準拠項目で同一のものであり，選択枝の内容もほぼ同じであるが，否定語を含むか否かによって正答率に大きな差が生じた (準拠 0.60，非準拠 0.35)。非準拠項目のコメント率は 2.2%と準拠項目の 0.7%に比べれば高いものの，多くの受検者にとって違和感のない項目であったと考えられる。つけられたコメントは「こうした項目の解き方を習った」という主旨のものであり，ストラテジーとして解答方略を知っているか否か

が正答／誤答を分けている可能性が示唆された。

同じ内容にも関わらず，大きく正答率に差が生じたことから，否定語を含む項目では認知負荷が高く，正確な読み取りが難しくなると考えられる。なお，コメントの中には，テストワイズネスに関わるものも存在し，否定語を含む選択枝において，テストワイズネスが影響している可能性も示唆された。これらのことから，本ガイドライン非準拠項目では，測定対象とした能力を十分に測定できていないことが示唆されたため，本ガイドラインは特に留意する必要があると考えられる。

ガイドライン 20e「明らかに不要・不自然な選択枝は入れないこと」

本ガイドラインの準拠項目では，説明されている古典文学作品を選択する項目であった。非準拠項目では，古典文学作品の選択枝の他，明らかに誤答とわかる現代漫画作品のタイトルを選択枝 4 として用いた。使用した項目を表 3-11 に示す。

表 3-11　ガイドライン 20e で用いた項目

20. 正答枝を探す手掛かりを与えないこと e. 明らかに不要・不自然な選択枝は入れないこと	
準拠	男性作者が女性になりきり執筆した，仮名文字を用いた日記形式の文学作品は何か。 1.　紫式部日記 2.　蜻蛉日記 3.* 土佐日記 4.　更級日記
非準拠	男性作者が女性になりきり執筆した，仮名文字を用いた日記形式の文学作品は何か。 1.　紫式部日記 2.　蜻蛉日記 3.* 土佐日記 4.　中学聖日記

非準拠項目について正答率は準拠項目よりも高くなっているものの，I-T 相関は低くなっている。非準拠項目では，正答枝のわからない受検者がランダムに選択をする際，3 枝選択になるのに対し，準拠項目では 4 枝選択になったためと考えられる。一方，非準拠項目において 5%ほどの受検者が明らかに誤答選択枝である選択枝 4 を選択した。つまり，誤答であるとわかっていながらも，あえてその選択枝を選択する者の存在が示唆された。

表 3-12 ガイドライン 13 で用いた項目

13.もっともらしく，識別力の高い選択枝のみにすること。多くの場合，3枝で事足りる	
準拠	以下の文の書き下し文として正しいものを選べ。 天帝使我長百獣。　戦国策・祖策 （天帝は私を多くの獣のかしらとした。） 1. 天帝百獣をして我に長たらしむ 2.* 天帝我をして百獣に長たらしむ 3. 天帝長をして我に百獣たらしむ
非準拠	以下の文の書き下し文として正しいものを選べ。 天帝使我長百獣。　戦国策・祖策 （天帝は私を多くの獣のかしらとした。） 1. 天帝百獣をして我に長たらしむ 2.* 天帝我をして百獣に長たらしむ 3. 天帝長をして我に百獣たらしむ 4. 天帝長をして百獣に我たらしむ

誤答枝として適切な第 4 枝がない場合には，ガイドライン 13「もっともらしく，識別力の高い選択枝のみにすること。多くの場合，3 枝で事足りる」(表 3-12)を参照し，3 枝選択項目とすることも考えられる。ガイドライン 13 では，非準拠項目において誤答とわかりやすい選択枝 4 を選択する受検者は少なく，正答率も準拠項目とほぼ変化がない。つまり，選択枝 4 は実質的に機能していない。Rodriguez (2005) では，メタ分析により選択枝数が 5 枝ないし 4 枝のものを 3 枝に減らしても，難易度や識別力には影響が少ないとしている。また，Shizuka, Takeuchi, Yashima, & Yoshizawa (2006) でも，3 枝選択式と 4 枝選択式の項目について，ほぼ難易度が変化しないことが示されている。つまり，項目の特性値という観点では，選択枝を 3 枝にすることに問題はない。しかし，ガイドライン 13 では準拠項目についてのみコメントがあった。内容は，テスト全体を通して選択枝数を見ると，この一題のみが 3 枝選択であることへの疑問を呈するものであった。その他の項目が全て 4 枝選択式であるため，本項目のみ 3 枝選択であることに違和感を覚えたようである。そのため，他の項目での出題形式や選択枝数を踏まえ，3 枝選択式にするか否かを検討する必要がある。

ガイドラインの総合的考察

ガイドラインの中には，そのガイドライン単体で検討すればよいもの (e.g.ガイドライン 4) と，ガイドラインを参照したとき，別のガイドラインについても同時に検討する必要性があるもの (e.g.ガイドライン 14 と 17) の存在が示唆された。複数のガイドラインを参照する必要があるものについては，同時にそれらのガイドラインに準拠すればいいということ

ではなく，項目の内容と受検者に与える印象を含めて検討する必要性が示唆された。

コメント率ならびにコメントの内容に着目すると，項目に違和感を覚えながらも項目自体がおかしいとは言わず，違和感について敢えてそのように出題されている理由を探す受検者もいた。ガイドライン 5 では，知識が不足しているため解答ができなかったというコメントがあった。ガイドライン 9 でも，全て平仮名で書かれた設問は障害への配慮だと考えた受検者もいた。つまり，項目へ違和感を覚えつつ，その違和感により解答が難しくなったとしても，テスト自体に間違いはないという考えが背後に見受けられた。作題する際はこうした点にも留意する必要があると考えられる。また，受検者にとってテストという存在の大きさを踏まえると，どのようなテスト項目を出題するかのみならず，そのテストの目的に適ったシステムになっているかどうか等，テストを取り巻くあらゆる環境についても十分に検討する必要性があるだろう。

3.5 考察

本研究では，同じ特性を測定する項目について，ガイドラインに準拠するか否かの影響を検討した。正答率や識別力へ影響を及ぼすガイドラインの存在が示され，その中には受検者にはほとんど気が付かれないものも存在した。

正答率・識別力および受検者の解答行動について着目した際に大きく影響を及ぼしていたガイドラインが，ガイドライン 18「『上記のいずれでもない』『上記すべてあてはまる』『分からない』などの選択枝を用いないこと」であった。準拠／非準拠によらず解答を導く際の傾向は同じであるにも関わらず，非準拠項目では，正答率・識別力が上がっていたことから，難易度は下がり解きやすい項目になっていた。

ガイドライン 3「各設問の内容は互いに独立であること」，ガイドライン 19「『でない』『〜以外』などの否定表現を用いないこと」では，準拠／非準拠により受検者の正答率に大きな差があった一方で，受検者のコメントは少なく，多くの受検者には気付かれないガイドラインであった。ガイドライン 5「解答が個人の意見に影響されないこと」では，準拠／非準拠による正答率の差は大きいものの，コメント率も高く，多くの受検者が気付くことができた。多くの受検者が気付くガイドラインとして，ガイドライン 4「重要な事項を問うこと．極端に細かかったり，逆に一般的すぎる内容にならないこと」，ガイドライン 9「言語レベルを受検者集団に合わせること」があるが，これらは正答率への影響は小さかった。ガイドライン 20e「明らかに不要・不自然な選択枝は入れないこと」では，非準拠では識別力が低くなる上に，機能しない誤答枝をあえて選択する受検者が存在するなどした。しかし，適切な誤答枝の作成が難しいとき，3 枝選択問題とするかは，その他の項目のバランスを踏まえて決定する必要がある。ガイドライン 13「もっともらしく，識別力の高い選択枝のみにすること。多くの場合，3 枝で事足りる」は，正答率への影響は小さく，コメントも少なかったものの，準拠項目にのみ違和感を覚えるコメントがあるガイドラインであった。

コメントの内容では，ガイドラインに非準拠である項目に対して，項目の違和感からコメントをする受検者がいる中で，違和感の理由として肯定的な解釈ができるものを探す受検者の存在も確認された。こうしたコメントをする受検者にとって，出題者は間違わないという意識が根底にあると考えられる。

本研究では，受検者にとって気が付かないガイドラインについて検討するために，コメント率を用いたが，気が付いていながらもコメントをしていない受検者も存在すると考えられる。更に，コメント率が低く，準拠／非準拠項目間で正答率が変わる項目についても，ストラテジーを用いて解答が可能であったというコメントが見られたように，そもそもストラテジーを知っているか否かが項目へ違和感を抱くか否かの背景となっている可能性がある。すなわち，ストラテジーを知らないから Flaw の存在に気が付かないのであり，知っていればそれを Flaw だと認識できる可能性が考えられる。この点において，本研究では分離

して検討することができない。

また，作成した項目は国語・数学・英語と複数の科目である一方で，教科特有の問題は検討していない。同じガイドラインであっても，教科が変われば影響が大きい／小さいという可能性がある。そのため，各教科固有の特徴を考慮した上で，それぞれの教科についてのガイドラインについても作成・検討されるべきであろう。

第4章

テストの評価に関する研究[4]

――*D* 指標のための群分けに関する研究（研究2）

[4] 本研究は，坪田・石井(印刷中)を改稿したものである。

4.1 問題と目的

本章では，テストの評価に関する指標についての検討を行う。*D* 指標は簡便な項目識別力の指標であり，受検者を群分けする必要がある。受検者の能力値(真値)による群分けから得られる *D* 指標の値により近くなるような受検者の群分け方法ならびにテスト条件について，クラスルームで行うテストを想定し，シミュレーションを用いた検討を行う。

項目識別力の指標のひとつである *D* 指標を算出する際には，受検者を得点の高低によって 3 群に分ける必要がある。3 群の分け方によっては，同じテストデータであっても算出される *D* 値が異なる。たとえば，カットオフポイント上にいる複数の同点者について同じ群として扱うのか，異なる群として扱うのかという分け方の違いによって，算出される *D* 指標の値は異なる。異なる群として扱う場合，テスト実施者が得られる正答数得点以外の情報を用いて受検者を群分けする必要がある。そのため，その他の情報として考えられる項目の正答率や識別力の指標をどのように用いて群分けをすることが望ましいのかを検討する必要がある。

また，クラスルームテストという条件下において，*D* 指標を用いることが実用上可能であるかについての検討はされていない。そのため *D* 指標を使用するために十分な信頼性を得るためのテスト条件を明らかにする必要がある。

以上のことをふまえ本研究では，より真値を反映した受検者の群分けの方法と望ましいテストの条件について，*D* 指標という側面から検討を行う。具体的には，下記の研究 2-1〜2-3 を行う。

研究 2-1 では，コンピュータシミュレーションによって，群分けの方法のひとつとして考えられる，カットオフポイント上の同点者を同じ群にした際のテストの信頼性ならびに受検者の順位づけについて評価する。

研究 2-2 では，カットオフポイント上の同点者を異なる群にするための並び替えの方法について，望ましいテスト条件と併せてコンピュータシミュレーションにより検討する。

研究 2-3 では，研究 2-1，2-2 の結果を踏まえ，実データの数値例を用いて，並び替え方法の有用性について確認する。

4.2 同点者を同じ群とする方法(研究 2-1)

目的

群を分ける方法として，同点の受検者を同じ群とすることが考えられる(e.g. 安永・斎藤・石井, 2010; 安永・石井, 2012)。同点を同じ群とすることで，上位群・中位群・下位群の比率は結果的に Kelly(1939)と異なるが，得点が同じ者が同じ群であることは直感的にわかりやすいと考えられる。ここでは，同点の受検者を同じ群とした際の *D* 指標への影響を，受検者数と項目数の観点から検討する。

方法

（1）群分けの方法

同点を同じ群として 3 群とするために考えられる以下 4 パタンで群分けを行った。

パタン1. 上位群と中位群ならびに中位群と下位群のカットオフポイントの受検者を全て中位群とする方法

パタン2. 上位群と中位群の間のカットオフポイントの受検者は中位群とし，中位群と下位群の間のカットオフポイントの受検者は下位群とする方法

パタン3. 上位群と中位群のカットオフポイントの受検者は上位群とし，中位群と下位群の間のカットオフポイントの受検者は中位群とする方法

パタン4. 上位群と中位群の間のカットオフポイントの受検者は上位群とし，中位群と下位群の間のカットオフポイントの受検者は下位群とする方法

（2）I-T 相関ならびに正答率の真値

実際に調査等で行われたテストの I-T 相関係数ならびに正答率が公表されている Azeem(2012)，Yu-mien(2010)，寺尾・安永・石井・野口(2015)を参考に，正答率ならびに I-T 相関の真値を，ベータ分布を用いて発生させた。

ベータ分布は 2 つの自由度の組み合わせにより分布が変化する確率分布である。$\beta(\alpha,\beta)$ のとき,

$$E(X)=\frac{\alpha}{(\alpha+\beta)}$$

$$V(X)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

となることが知られている。

正答率ならびに I-T 相関係数について, 具体的にはそれぞれ, 以下のベータ分布を用いた。正答率は，一様分布となる $\beta(1,1)$ (M=0.5,SD=0.289)，低い正答率で固まる場合 $\beta(2,6)$ (M=0.25,SD=0.144), 中程度の正答率で固まる場合 $\beta(6,6)$ (M=0.5,SD=0.139), 高い正答率で固

まる場合$\beta(6,2)$ (M=0.75,SD=0.144)の 4 パタンについて検討した。I-T 相関は，低い識別力の場合$\beta(1,9)$ (M=0.10,SD=0.090)，中程度の識別力の場合$\beta(4,12)$ (M=0.25,SD=0.105)，高い識別力の場合$\beta(10,15)$ (M=0.40,SD=0.096)の 3 パタンについて検討した。

これらの組み合わせ計 12 パタンについて検討を行った。

（3）困難度・識別力の変換の方法

シミュレーションは，2 Parameter Logistic Model (2PLM)を用いた。このモデルを用いることで，受検者の能力値の分布について，標準正規分布を真値として仮定することが可能となる。

2PLM では，項目の困難度と識別力は潜在特性曲線を描くパラメタとなる。そのため，（2）で述べた項目の正答率ならびに I-T 相関係数の真値について，Lord &Novick(1968), 松井(1991) を参考に，項目 j のときの I-T 相関係数ρ_jと，標準正規分布の上側確率が正答率となる z 値γ_jを用いて，以下のように変換を行った。

$$a_j = \frac{\rho_j}{\sqrt{1-\rho_j}}$$

$$b_j = \frac{\gamma_j}{\rho_j}$$

（4）テストの条件

テストの条件として，受検者数と項目数を設定した。受検者数は N=10, 40, 70, 100, 200 の 5 パタン，項目数は m=10, 20, 30, 40, 50 の 5 パタンとした．これらの組み合わせ 25 パタンについて検討を行った。

（5）シミュレーションの手続き

以下の手続きによりシミュレーションを行った。

1. 項目数の正答率ならびに識別力の真値を発生させ，これらを 2 PLM の識別力と困難度に対応するように変換を行った。
2. 受検者の潜在特性真値を$N(0,1)$から受検者数分発生させ，1 で求めた項目特性値を用い，各項目の正答確率を求めた。
3. 受検者ごとに一様乱数を発生させ，2 で求めた正答確率を用いて，正誤パタンを作成した。
4. 合計得点順に並び替え，パタン 1〜4 の方法で群分けを行った。
5. 群分けに基づき，真値との比較として，重み付けカッパ係数ならびに ΔD を算出した。
6. 以上の手続きを全ての受検者数・項目数・正答率・I-T 相関の組み合わせについて 100 回ずつ行った。

（6）用いる指標

真値に基づいた群分けと，提案手法による群分けの比較をするために，重み付けカッパ係数，ΔD という 2 つの指標を用いた。

重み付けカッパ係数は，真値と提案手法の群分けの一致度についての指標である。真値による群分けの上位群・中位群・下位群と，提案手法による群分けの上位群・中位群・下位群という 3 群のクロス表について，それぞれのセルに含まれるケース数を用いることで一致度合いについて検討するものであり，下記式により算出される。

$$k = 1 - \frac{\Sigma_{i,j} w_{i,j} x_{i,j}}{\Sigma_{i,j} w_{i,j} m_{i,j}}$$

$w_{i,j}$は，真値による群分けと提案手法による群分けについての重みであり，$(i-j)^2$で表される。つまり，判断された群が離れているほど，重みが大きくなる。たとえば，真値による群分けでは上位群であり，提案手法による群分けでは下位群となったケースでは，$(1-3)^2 = 4$がその重みとして算出される。

$x_{i,j}$は真値による群分けと提案手法による群分けについて，各セルに含まれるケース数から観測割合を算出したものである。たとえば，真値による群分けでは上位群であり，提案手法による群分けでは下位群となったケースが全体の何割であったかを考える。

$m_{i,j}$は各セルの期待値を表す。たとえば，真値による群分けでは上位群であった割合と，提案手法による群分けでは下位群となった割合を掛けることで期待値を算出することができる。

重み付けカッパ係数は，真値による群分けと提案手法の群分けがどの程度一致しているかどうかを表す指標である。真値による群分けと提案手法の群分けが完全に一致していたとき，k=1 となり，一致の程度が低くなるについて，0 に近づく。

D 指標の値は下記式により算出される。

$$D = \bar{X}_{upper} - \bar{X}_{lower}$$

$\bar{X}_{upper}$は上位群のテストの平均正答率を表し，$\bar{X}_{lower}$は下位群のテストの平均正答率を表す。

ΔD は，真値による群分けにより算出された D をD_Tと，提案手法による群分けにより算出された D をD_Pとすると，以下のように表される。

$$\Delta D = D_T - D_P$$

ΔD は，真値の群分けにより算出された D 指標と提案手法の群分けにより算出された D 指標の値の差である。つまり，群分けの方法によって，テストの識別力の値が真の D からどの程度の差を表すかを検討するためのものであり，真値と提案手法の値が近くなるほど 0 に近づく。

結果と考察

重みづけカッパ係数ならびに ΔD の結果の一部を図 4-1〜4-4 に示した。図 4-1,4-2 は識別力が低く困難度が一様である条件，図 4-3,4-4 は識別力が高く困難度が一様である条件である。重みづけカッパ係数は 1.0 に近いほど，真値と推定値の順序が一致していることになる。

また，ΔD は 0.0 に近いほど真値と推定値の D 値の差が小さいことになる。

低識別力よりも高識別力で重みづけカッパ係数は大きい値をとり，ΔD は 0 に近い値をとった。また，識別力によらず，項目数が大きくなるにつれ，重みづけカッパ係数は大きい値をとり，ΔD は 0 に近い値をとった。

群分けを行ったパタンのうち，パタン 4 以外の方法では条件によらず，上位群もしくは下位群の該当者がいないことで，そもそも D 指標の算出が不可能であるケースが見られた。パタン 1〜3 ではカットオフポイント上の受検者を中位群とするため，カットオフポイント以上/以下の受検者が居ないとき，上位群/下位群に入る受検者が存在しないということになる。

重み付けカッパ係数の値を大きくし，ΔD を小さくする要因として，人数はほとんど影響しないと考えられる。図 4-1〜4-4 のいずれの条件であっても，受検者数が 40 名以上ではグラフの傾きがほぼ横ばいとなった。すなわち，テストを評価するための受検者数の最低単位は，クラスルームテストにおける 1 クラスの生徒数だと考えられる 40 名程度だと考えられる。

一方，真値による群分けに近づけるような，重み付けカッパ係数の値が高くなり，ΔD の値が小さくなる要因として，テストの識別力条件と項目数が大きく影響すると考えられる。テスト項目の識別力が高く，かつ項目数が多ければ多いほど，重み付けカッパ係数の値は大きくなり，ΔD の値は小さくなった。しかしながら，クラスルームテストでの項目数を考えたとき，50 項目を超えるテストは少ない。更に，本シミュレーションの識別力条件は，実際のテストで算出される識別力よりも概して大きいと考えられる。

ΔD は，識別力の条件や項目数による影響が小さい。重み付けカッパ係数と合わせて考えると，識別力が低いテストでは，重み付けカッパ係数が低いことから，受検者の群分けがランダムに近いものの，そもそもの真値による D 指標自体が小さい値であることで真値との差が小さくなったと考えられる。対して，識別力が高いテストでは，識別力が低いテストよりも重み付けカッパ係数が高いことから，より真値に近い群分けになり，真値との差が小さくなったと考えられる。

このように，ΔD を 0 に近づけること以上に，重み付けカッパ係数の値が大きくなる条件を検討する必要がある。また，同点者を同じ群とすることで，該当者が 0 人となる群分けが行われることは問題である。そのため研究 2-2 では，合計得点が同じ受検者を異なる群に分けることで，また，合計得点が同じ受検者を異なる群に分ける手法によって，どのような条件で重み付けカッパ係数の値が大きくなるかを検討する。

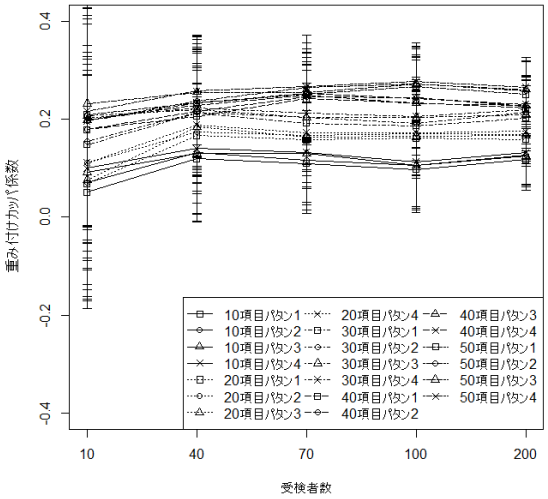

図 4-1　低識別力・困難度一様の
重みづけカッパ係数

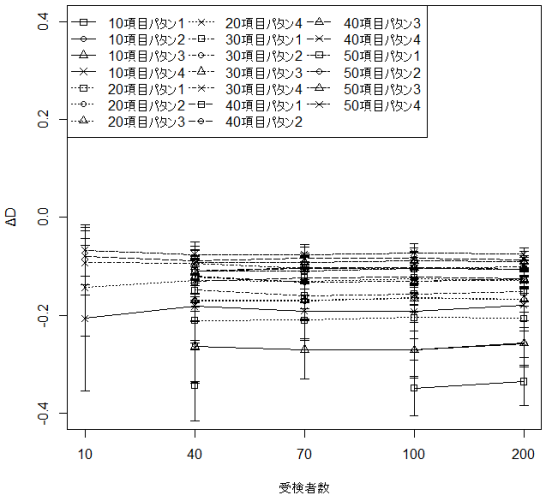

図 4-2　低識別力・困難度一様の ΔD

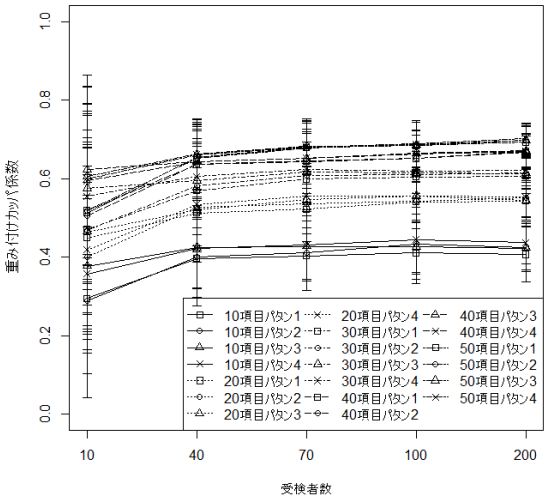

図 4-3　高識別力・困難度一様の
重みづけカッパ係数

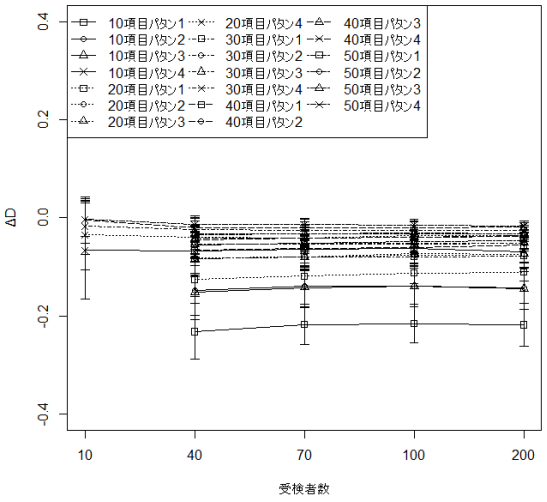

図 4-4　高識別力・困難度一様の ΔD

4.3 同点者を異なる群とする方法(研究 2-2)

問題と目的

前節では，上位群と中位群，中位群と下位群のカットオフポイント上にいる同じ得点の受検者について，同じ群となるように群分けを行い，*D* 指標ならびに重み付けカッパ係数について検討を行った。しかし，複数の条件で上位群，中位群，下位群のいずれかの群に属する受検者が存在せず，そもそも *D* 指標の算出が不可能となる場合があった。また，より真の群分けの状態に近づけるためにも，重み付けカッパ係数を大きくすることが求められる。そのため，カットオフポイント上にいる同点の受検者について，異なる群に分ける方法について検討が必要である。

そこで本節では，上位群と中位群，中位群と下位群のカットオフポイント上にいる同じ得点の受検者について，Kelly(1939)の用いた割合に一致するよう，異なる群に割り当てる方法について検討する。

同点者を異なる群に割り当てるにあたり，合計得点以外の情報を用いて受検者の順位付けを行う必要が生じる。そこで，より真値の並び順と一致する方法について，識別力，正答率を用いて，探索的に検討を行う。その上で，より真値に近い並び順を可能とするテスト条件についても検討する。

方法

（1）並び変える方法

まず，合計得点を降順に並び替える。合計得点が同点だった場合の受検者を並び替える方法として，ID(受検者番号)，I-T 相関，正答率を用いた以下の 9 パタンについて検討した。

Type1.　ID の大小について，ID 番号順に並び替える。このとき，ID はランダムに振り分けられている番号である

Type2.1. 項目内で最も I-T 相関が高かった項目への正誤について降順に並び替える

Type2.2. 項目内で最も I-T 相関が高かった項目への正誤，次に I-T 相関が高かった項目への正誤について降順に並び替える

Type2.3. 項目内で最も I-T 相関が高かった項目への正誤，次に I-T 相関が高かった項目への正誤，さらにその次に I-T 相関が高かった項目への正誤について降順に並び替える

Type2.4. 受検者が正答した項目のうち，I-T 相関がより高い項目に正答していた受検者を上位として降順に並び替える

Type3.1. 正答率が 0.5 に最も近い項目への正誤について降順に並び替える

Type3.2. 正答率が上位群と中位群のカットオフポイントに最も近い項目への正誤，中位群と下位群のカットオフポイントに最も近い項目への正誤について降順に並び替える

Type3.3. 正答率が 0.5 に最も近い項目への正誤，上位群と中位群のカットオフポイントに最も近い項目への正誤，中位群と下位群のカットオフポイントに最も近い項目への正誤について降順に並び替える

Type3.4. 受検者が正答した項目のうち，正答率がより高い項目に正答していた受検者を上位として降順に並び替える

（2）I-T 相関ならびに正答率の真値

設定は研究 2-1 で用いたものと同じである。

（3）識別力・困難度の変換の方法

研究 2-1 と同一の方法により変換を行った。

（4）テストの条件

研究 2-1 で得られた条件を踏まえ，テストの受検者数は N=40, 100 の 2 パタン，テストの項目数は m=10, 30, 50 の 3 パタンとした。これらの組み合わせ 6 パタンについて検討を行った。

（5）シミュレーションの手続き

以下の手続きによりシミュレーションを行った。

1. 項目数の正答率ならびに識別力の真値を発生させ，これらを 2 PLM の識別力と困難度に対応するように変換を行った。
2. 受検者の潜在特性真値を$N(0,1)$から受検者数分発生させ，1 で求めた項目特性値を用い，各項目の正答確率を求めた。
3. 受検者ごとに一様乱数を発生させ，2 で求めた正答確率を用いて，正誤パタンを作成した。
4. 正誤パタンより，各項目の正答率ならびに I-T 相関を算出した。
5. Type1〜3.4 の方法で並び替え，それに基づき群分けを行った。
6. 群分けに基づき，真値との比較として，順位相関係数，重み付けカッパ係数ならびに ΔD を算出した。

以上の手続きを全ての受検者数・項目数・正答率・I-T 相関の組み合わせについて 100 回ずつ行った。

（6）用いる指標

真値と提案手法による並び順ならびに群分けの比較をするために，スピアマンの順位相関係数，重み付けカッパ係数を用いた。また，真値による群分けの識別力と，テスト条件による群分けの識別力を比較するために，ΔD を用いた。

スピアマンの順位相関係数は，真値の並び順と提案手法の並び順がどの程度一致しているかという指標である。受検者数を n とおいたとき，スピアマンの順位相関 r は下記の式により算出される。

$$r = \frac{\sum_{i=1}^{n} R_n}{n(n-1)}$$

このとき，R_n は真値の並び順と提案手法の並び順による順位の差を表す。

また，重み付けカッパ係数と ΔD は，研究 2-1 と同じ指標である。

結果と考察

順位相関係数，重みづけカッパ係数ならびに ΔD が，表 4-1～4-24 である。ここで検討している並び替えの方法は，同点の受検者についての並び替えであり，上位群，中位群，下位群のそれぞれの平均点はどのような並び替えを行っても同じになる。そのため，ΔD は各条件につき一つ算出された。

項目数が増えるほど，順位相関係数ならびに重み付けカッパ係数の値は大きくなった。識別力条件について検討すると，識別力が高くなるにつれて順位相関係数，重み付けカッパ係数の値は大きくなる。また，困難度条件について確認すると，最も順位相関係数ならびに重み付けカッパ係数の値を最も大きくするのは，困難度が中程度付近で固まる条件であり，次いで一様分布の条件であった。

Lord(1982) は，平行テスト x, y におけるの信頼性の基準として，$\rho_{xy} = .90$ としている。真値と並び替え順の順位相関係数の値を代用すると，その基準を満たす重みづけカッパ係数はおよそ 0.7 を示していることより，信頼できる D 指標を算出するための重みづけカッパ係数の値は，最低 0.7 程度であることが考えられる。これを踏まえテスト条件を確認すると，識別力が高く，50 項目程度の項目数があることが望ましい。しかし，30 項目程度以上でも識別力が高く困難度が中程度に固まる場合などの特定の条件によっては，かなり高い r の値を得た。30 項目という項目数は，クラスルームテストとして実施するテストでは，妥当な数であると考えられる。

また，同点者を並び替える手法としては，多くのテスト条件で Type2.1～2.4 の識別力を用いる手法が望ましいことが示唆された。特に，Type2.4 のように，受検者の全ての正答項目の識別力を用いた方法がより良い。一方，実際のテスト現場での実用性を鑑みると，識別力の高い数項目について並び替えを行うだけでも，ランダムである ID 順に並び替えるよりは望ましい結果となることが示された。

また，識別力が高い条件では，並び替えの影響がほとんど見られず，どのような手法で並び替えても重み付けカッパ係数の値は一貫して大きかった。元々識別力の高い項目で作成されたテストでは，合計得点が算出された時点で十分に受検者の能力の高低を識別できているため，識別力による並び替えの影響は小さいのだと考えられる。

ΔD を見ると，識別力が大きくなるにつれて，また，項目数が大きくなるにつれて 0 に近い値となった。実際のテスト場面において，識別力の検討はテスト実施後であることを考慮すると，項目分析のためには最低でも 30 項目程度が必要になると考えられる。

以上のことより，項目分析を行うためには，テストの項目数は 30 項目以上であり，識別

力が高いテストを作成することが望ましい。また，合計得点が同じ受検者について，識別力の高い項目への正誤を用いて並べ替えを行うことで，より真値に近い順位になることがわかった。

表 4-1　低識別力・困難度一様の各指標(N=40)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.254 (0.067)	r	0.186 (0.170)	0.185 (0.175)	0.185 (0.177)	0.185 (0.174)	0.189 (0.173)	0.189 (0.176)	0.187 (0.174)	0.189 (0.174)	0.187 (0.167)
		k	0.107 (0.133)	0.106 (0.139)	0.112 (0.139)	0.110 (0.141)	0.112 (0.134)	0.107 (0.136)	0.105 (0.132)	0.107 (0.130)	0.119 (0.133)
30	-0.124 (0.030)	r	0.338 (0.139)	0.339 (0.143)	0.339 (0.143)	0.341 (0.143)	0.347 (0.141)	0.336 (0.144)	0.336 (0.144)	0.334 (0.143)	0.338 (0.139)
		k	0.216 (0.119)	0.219 (0.122)	0.222 (0.119)	0.223 (0.119)	0.221 (0.113)	0.216 (0.119)	0.215 (0.123)	0.215 (0.121)	0.220 (0.113)
50	-0.085 (0.022)	r	0.409 (0.138)	0.410 (0.137)	0.412 (0.137)	0.412 (0.138)	0.417 (0.135)	0.415 (0.139)	0.415 (0.137)	0.418 (0.137)	0.413 (0.133)
		k	0.255 (0.136)	0.257 (0.135)	0.263 (0.133)	0.265 (0.133)	0.265 (0.128)	0.255 (0.136)	0.253 (0.134)	0.258 (0.132)	0.257 (0.140)

() 内は標準偏差を表す

表 4-2　低識別力・低困難度の各指標(N=40)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.264 (0.062)	r	0.172 (0.190)	0.173 (0.186)	0.169 (0.186)	0.170 (0.186)	0.172 (0.187)	0.175 (0.186)	0.170 (0.185)	0.172 (0.185)	0.170 (0.183)
		k	0.101 (0.130)	0.105 (0.132)	0.095 (0.130)	0.099 (0.135)	0.109 (0.136)	0.094 (0.134)	0.099 (0.132)	0.100 (0.138)	0.103 (0.139)
30	-0.123 (0.038)	r	0.351 (0.157)	0.351 (0.160)	0.352 (0.159)	0.351 (0.160)	0.349 (0.166)	0.348 (0.158)	0.347 (0.164)	0.348 (0.161)	0.345 (0.165)
		k	0.221 (0.131)	0.221 (0.132)	0.223 (0.135)	0.218 (0.132)	0.221 (0.133)	0.219 (0.140)	0.220 (0.134)	0.221 (0.139)	0.215 (0.142)
50	-0.086 (0.022)	r	0.437 (0.135)	0.437 (0.136)	0.438 (0.137)	0.438 (0.137)	0.441 (0.133)	0.438 (0.135)	0.437 (0.134)	0.439 (0.134)	0.437 (0.132)
		k	0.257 (0.117)	0.255 (0.118)	0.261 (0.121)	0.262 (0.120)	0.263 (0.125)	0.262 (0.126)	0.258 (0.124)	0.264 (0.125)	0.264 (0.123)

() 内は標準偏差を表す

表 4-3　低識別力・中困難度の各指標(N=40)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.288 (0.071)	r	0.238 (0.148)	0.237 (0.146)	0.238 (0.148)	0.236 (0.148)	0.233 (0.147)	0.238 (0.146)	0.237 (0.142)	0.238 (0.142)	0.238 (0.139)
		k	0.139 (0.129)	0.131 (0.129)	0.133 (0.127)	0.134 (0.125)	0.143 (0.126)	0.139 (0.135)	0.141 (0.127)	0.141 (0.129)	0.142 (0.131)
30	-0.130 (0.042)	r	0.413 (0.158)	0.415 (0.158)	0.417 (0.159)	0.419 (0.158)	0.418 (0.157)	0.414 (0.155)	0.415 (0.157)	0.416 (0.156)	0.409 (0.155)
		k	0.263 (0.141)	0.260 (0.143)	0.262 (0.146)	0.263 (0.148)	0.259 (0.142)	0.260 (0.139)	0.256 (0.146)	0.257 (0.142)	0.255 (0.150)
50	-0.095 (0.024)	r	0.474 (0.126)	0.476 (0.126)	0.477 (0.125)	0.477 (0.124)	0.478 (0.124)	0.473 (0.126)	0.472 (0.123)	0.472 (0.125)	0.472 (0.126)
		k	0.293 (0.123)	0.296 (0.121)	0.295 (0.119)	0.297 (0.120)	0.293 (0.119)	0.289 (0.121)	0.283 (0.120)	0.289 (0.122)	0.287 (0.122)

() 内は標準偏差を表す

表 4-4　低識別力・高困難度の各指標(N=40)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.252 (0.066)	r	0.187 (0.153)	0.191 (0.158)	0.192 (0.159)	0.192 (0.159)	0.191 (0.159)	0.191 (0.158)	0.184 (0.147)	0.187 (0.156)	0.185 (0.152)
		k	0.115 (0.132)	0.112 (0.129)	0.112 (0.129)	0.107 (0.132)	0.104 (0.126)	0.108 (0.133)	0.108 (0.129)	0.100 (0.138)	0.107 (0.128)
30	-0.124 (0.036)	r	0.327 (0.151)	0.328 (0.152)	0.330 (0.154)	0.333 (0.153)	0.336 (0.152)	0.330 (0.152)	0.323 (0.157)	0.327 (0.153)	0.325 (0.154)
		k	0.197 (0.121)	0.201 (0.121)	0.204 (0.122)	0.207 (0.122)	0.205 (0.126)	0.202 (0.122)	0.203 (0.128)	0.203 (0.126)	0.202 (0.118)
50	-0.089 (0.024)	r	0.397 (0.138)	0.398 (0.138)	0.397 (0.137)	0.396 (0.138)	0.399 (0.144)	0.396 (0.142)	0.397 (0.141)	0.398 (0.143)	0.399 (0.139)
		k	0.245 (0.130)	0.247 (0.127)	0.246 (0.124)	0.245 (0.122)	0.259 (0.128)	0.253 (0.127)	0.251 (0.125)	0.253 (0.129)	0.255 (0.116)

() 内は標準偏差を表す

表 4-5　中識別力・困難度一様の各指標(N=40)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.182 (0.053)	r	0.441 (0.151)	0.444 (0.147)	0.446 (0.143)	0.447 (0.145)	0.448 (0.148)	0.439 (0.152)	0.435 (0.148)	0.436 (0.149)	0.437 (0.151)
		k	0.273 (0.145)	0.277 (0.131)	0.282 (0.134)	0.283 (0.129)	0.276 (0.135)	0.271 (0.139)	0.269 (0.131)	0.265 (0.137)	0.276 (0.124)
30	-0.076 (0.023)	r	0.664 (0.091)	0.667 (0.089)	0.670 (0.089)	0.669 (0.090)	0.671 (0.091)	0.662 (0.094)	0.666 (0.094)	0.665 (0.094)	0.664 (0.094)
		k	0.454 (0.105)	0.448 (0.102)	0.456 (0.105)	0.450 (0.116)	0.459 (0.111)	0.449 (0.113)	0.463 (0.117)	0.463 (0.112)	0.454 (0.112)
50	-0.050 (0.017)	r	0.745 (0.089)	0.747 (0.090)	0.748 (0.090)	0.747 (0.089)	0.751 (0.088)	0.747 (0.089)	0.745 (0.088)	0.745 (0.089)	0.747 (0.089)
		k	0.523 (0.110)	0.523 (0.111)	0.522 (0.112)	0.521 (0.111)	0.523 (0.106)	0.523 (0.111)	0.523 (0.112)	0.522 (0.111)	0.526 (0.113)

() 内は標準偏差を表す

表 4-6　中識別力・低困難度の各指標(N=40)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.178 (0.051)	r	0.437 (0.131)	0.442 (0.131)	0.444 (0.133)	0.446 (0.133)	0.448 (0.132)	0.435 (0.130)	0.440 (0.131)	0.438 (0.131)	0.441 (0.135)
		k	0.291 (0.117)	0.290 (0.116)	0.291 (0.118)	0.295 (0.119)	0.283 (0.114)	0.279 (0.118)	0.290 (0.121)	0.285 (0.121)	0.286 (0.127)
30	-0.071 (0.025)	r	0.681 (0.106)	0.681 (0.107)	0.682 (0.106)	0.682 (0.106)	0.684 (0.105)	0.682 (0.105)	0.683 (0.109)	0.682 (0.108)	0.679 (0.107)
		k	0.473 (0.130)	0.469 (0.134)	0.471 (0.132)	0.465 (0.138)	0.472 (0.131)	0.474 (0.129)	0.466 (0.132)	0.463 (0.131)	0.459 (0.133)
50	-0.051 (0.018)	r	0.753 (0.090)	0.755 (0.089)	0.755 (0.088)	0.755 (0.090)	0.758 (0.089)	0.753 (0.090)	0.754 (0.090)	0.754 (0.091)	0.753 (0.092)
		k	0.535 (0.121)	0.533 (0.119)	0.538 (0.120)	0.535 (0.121)	0.540 (0.121)	0.529 (0.118)	0.537 (0.122)	0.535 (0.119)	0.533 (0.121)

() 内は標準偏差を表す

表 4-7　中識別力・中困難度の各指標(N=40)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.194 (0.063)	r	0.510 (0.153)	0.513 (0.150)	0.515 (0.150)	0.514 (0.153)	0.517 (0.150)	0.510 (0.149)	0.514 (0.148)	0.513 (0.146)	0.511 (0.152)
		k	0.337 (0.135)	0.343 (0.135)	0.339 (0.139)	0.336 (0.144)	0.327 (0.146)	0.336 (0.126)	0.331 (0.136)	0.330 (0.133)	0.337 (0.140)
30	-0.074 (0.024)	r	0.731 (0.081)	0.734 (0.080)	0.736 (0.079)	0.737 (0.079)	0.739 (0.079)	0.730 (0.084)	0.732 (0.083)	0.731 (0.085)	0.734 (0.082)
		k	0.511 (0.121)	0.518 (0.110)	0.519 (0.110)	0.524 (0.111)	0.523 (0.120)	0.508 (0.119)	0.513 (0.116)	0.511 (0.116)	0.515 (0.117)
50	-0.050 (0.019)	r	0.796 (0.062)	0.799 (0.062)	0.797 (0.063)	0.798 (0.063)	0.800 (0.064)	0.796 (0.063)	0.797 (0.063)	0.797 (0.063)	0.796 (0.064)
		k	0.584 (0.098)	0.589 (0.099)	0.589 (0.103)	0.591 (0.104)	0.594 (0.098)	0.586 (0.103)	0.593 (0.097)	0.595 (0.099)	0.594 (0.098)

() 内は標準偏差を表す

表 4-8　中識別力・高困難度の各指標(N=40)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.175 (0.059)	r	0.450 (0.144)	0.459 (0.144)	0.463 (0.142)	0.464 (0.143)	0.461 (0.145)	0.452 (0.142)	0.450 (0.141)	0.450 (0.140)	0.450 (0.135)
		k	0.294 (0.146)	0.304 (0.147)	0.306 (0.144)	0.309 (0.144)	0.301 (0.140)	0.287 (0.144)	0.291 (0.147)	0.284 (0.150)	0.292 (0.134)
30	-0.073 (0.025)	r	0.667 (0.101)	0.669 (0.102)	0.669 (0.102)	0.669 (0.102)	0.674 (0.101)	0.667 (0.100)	0.664 (0.102)	0.666 (0.101)	0.664 (0.100)
		k	0.463 (0.116)	0.460 (0.119)	0.461 (0.116)	0.464 (0.117)	0.467 (0.126)	0.461 (0.121)	0.449 (0.121)	0.457 (0.123)	0.461 (0.125)
50	-0.052 (0.017)	r	0.735 (0.082)	0.735 (0.081)	0.735 (0.081)	0.736 (0.081)	0.739 (0.082)	0.735 (0.082)	0.735 (0.083)	0.735 (0.083)	0.733 (0.083)
		k	0.528 (0.113)	0.527 (0.111)	0.528 (0.113)	0.525 (0.110)	0.527 (0.111)	0.521 (0.115)	0.519 (0.114)	0.516 (0.115)	0.519 (0.114)

() 内は標準偏差を表す

表 4-9　高識別力・困難度一様の各指標(N=40)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.129 (0.046)	r	0.645 (0.109)	0.647 (0.110)	0.646 (0.110)	0.649 (0.110)	0.648 (0.106)	0.647 (0.110)	0.648 (0.110)	0.647 (0.110)	0.645 (0.106)
		k	0.437 (0.128)	0.432 (0.126)	0.432 (0.125)	0.435 (0.129)	0.429 (0.134)	0.429 (0.132)	0.433 (0.136)	0.430 (0.134)	0.432 (0.130)
30	-0.047 (0.016)	r	0.826 (0.059)	0.827 (0.059)	0.827 (0.059)	0.828 (0.059)	0.828 (0.059)	0.826 (0.059)	0.825 (0.060)	0.825 (0.059)	0.825 (0.060)
		k	0.632 (0.094)	0.631 (0.090)	0.626 (0.090)	0.625 (0.091)	0.631 (0.095)	0.627 (0.091)	0.625 (0.092)	0.625 (0.094)	0.627 (0.094)
50	-0.031 (0.016)	r	0.880 (0.045)	0.880 (0.045)	0.880 (0.045)	0.880 (0.046)	0.882 (0.045)	0.880 (0.043)	0.881 (0.045)	0.880 (0.044)	0.879 (0.045)
		k	0.683 (0.104)	0.679 (0.106)	0.676 (0.107)	0.677 (0.108)	0.686 (0.104)	0.683 (0.103)	0.681 (0.110)	0.685 (0.107)	0.683 (0.101)

() 内は標準偏差を表す

表 4-10　高識別力・低困難度の各指標(N=40)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.120 (0.038)	r	0.651 (0.089)	0.653 (0.089)	0.654 (0.091)	0.654 (0.090)	0.656 (0.092)	0.651 (0.098)	0.650 (0.092)	0.648 (0.095)	0.651 (0.092)
		k	0.453 (0.102)	0.454 (0.103)	0.453 (0.102)	0.447 (0.110)	0.455 (0.106)	0.446 (0.107)	0.452 (0.106)	0.447 (0.105)	0.451 (0.109)
30	-0.050 (0.018)	r	0.815 (0.063)	0.815 (0.063)	0.815 (0.063)	0.815 (0.064)	0.817 (0.064)	0.815 (0.063)	0.814 (0.063)	0.815 (0.063)	0.816 (0.062)
		k	0.611 (0.102)	0.611 (0.101)	0.611 (0.102)	0.612 (0.102)	0.611 (0.108)	0.607 (0.104)	0.610 (0.107)	0.612 (0.105)	0.612 (0.106)
50	-0.030 (0.014)	r	0.872 (0.043)	0.873 (0.044)	0.874 (0.044)	0.874 (0.044)	0.875 (0.042)	0.873 (0.043)	0.872 (0.044)	0.873 (0.043)	0.873 (0.043)
		k	0.683 (0.092)	0.683 (0.092)	0.685 (0.092)	0.685 (0.093)	0.686 (0.092)	0.685 (0.089)	0.681 (0.098)	0.684 (0.095)	0.687 (0.091)

() 内は標準偏差を表す

表 4-11　高識別力・中困難度の各指標(N=40)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.131 (0.046)	r	0.711 (0.084)	0.713 (0.085)	0.714 (0.086)	0.717 (0.085)	0.716 (0.084)	0.711 (0.083)	0.711 (0.086)	0.712 (0.084)	0.716 (0.083)
		k	0.501 (0.103)	0.501 (0.108)	0.501 (0.110)	0.500 (0.109)	0.499 (0.106)	0.497 (0.111)	0.499 (0.114)	0.490 (0.111)	0.503 (0.110)
30	-0.054 (0.025)	r	0.866 (0.050)	0.866 (0.051)	0.866 (0.050)	0.867 (0.049)	0.868 (0.048)	0.865 (0.049)	0.866 (0.049)	0.865 (0.049)	0.866 (0.049)
		k	0.662 (0.103)	0.663 (0.108)	0.662 (0.106)	0.659 (0.110)	0.660 (0.106)	0.658 (0.104)	0.662 (0.106)	0.659 (0.103)	0.663 (0.105)
50	-0.031 (0.013)	r	0.908 (0.033)	0.908 (0.033)	0.908 (0.033)	0.908 (0.033)	0.908 (0.034)	0.908 (0.033)	0.908 (0.033)	0.908 (0.034)	0.908 (0.033)
		k	0.739 (0.080)	0.737 (0.080)	0.734 (0.081)	0.731 (0.079)	0.733 (0.082)	0.737 (0.079)	0.735 (0.079)	0.733 (0.080)	0.733 (0.078)

() 内は標準偏差を表す

表 4-12　高識別力・高困難度の各指標(N=40)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.130 (0.045)	r	0.635 (0.109)	0.637 (0.109)	0.638 (0.109)	0.641 (0.109)	0.644 (0.109)	0.636 (0.112)	0.637 (0.110)	0.638 (0.110)	0.636 (0.110)
		k	0.421 (0.116)	0.426 (0.115)	0.429 (0.109)	0.431 (0.109)	0.435 (0.114)	0.419 (0.123)	0.421 (0.118)	0.421 (0.121)	0.425 (0.116)
30	-0.053 (0.021)	r	0.812 (0.055)	0.811 (0.054)	0.813 (0.054)	0.813 (0.054)	0.817 (0.051)	0.812 (0.054)	0.813 (0.055)	0.813 (0.054)	0.811 (0.057)
		k	0.591 (0.123)	0.589 (0.122)	0.589 (0.120)	0.588 (0.126)	0.594 (0.125)	0.591 (0.123)	0.591 (0.123)	0.591 (0.118)	0.586 (0.117)
50	-0.031 (0.013)	r	0.876 (0.043)	0.877 (0.042)	0.878 (0.042)	0.878 (0.042)	0.878 (0.043)	0.878 (0.042)	0.877 (0.041)	0.878 (0.042)	0.875 (0.044)
		k	0.673 (0.099)	0.671 (0.099)	0.672 (0.100)	0.673 (0.104)	0.674 (0.095)	0.677 (0.095)	0.668 (0.096)	0.671 (0.092)	0.672 (0.097)

() 内は標準偏差を表す

表 4-13　低識別力・困難度一様の各指標(N=100)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.253 (0.043)	r	0.193 (0.109)	0.193 (0.108)	0.193 (0.109)	0.191 (0.111)	0.190 (0.114)	0.194 (0.113)	0.192 (0.112)	0.194 (0.113)	0.184 (0.111)
		k	0.116 (0.078)	0.114 (0.082)	0.115 (0.079)	0.112 (0.084)	0.113 (0.085)	0.115 (0.084)	0.117 (0.083)	0.119 (0.086)	0.109 (0.080)
30	-0.118 (0.022)	r	0.362 (0.105)	0.366 (0.107)	0.368 (0.107)	0.368 (0.108)	0.369 (0.107)	0.360 (0.106)	0.361 (0.106)	0.361 (0.106)	0.361 (0.106)
		k	0.233 (0.091)	0.236 (0.095)	0.234 (0.092)	0.237 (0.091)	0.239 (0.094)	0.226 (0.092)	0.232 (0.087)	0.228 (0.093)	0.231 (0.092)
50	-0.086 (0.017)	r	0.427 (0.105)	0.430 (0.107)	0.432 (0.106)	0.432 (0.107)	0.435 (0.108)	0.426 (0.106)	0.426 (0.105)	0.426 (0.105)	0.427 (0.104)
		k	0.273 (0.090)	0.277 (0.096)	0.281 (0.093)	0.283 (0.095)	0.285 (0.095)	0.273 (0.094)	0.278 (0.093)	0.277 (0.092)	0.277 (0.090)

() 内は標準偏差を表す

表 4-14　低識別力・低困難度の各指標(N=100)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.240 (0.039)	r	0.188 (0.117)	0.189 (0.120)	0.192 (0.120)	0.192 (0.120)	0.192 (0.121)	0.188 (0.120)	0.191 (0.116)	0.190 (0.120)	0.189 (0.119)
		k	0.112 (0.087)	0.115 (0.089)	0.113 (0.086)	0.114 (0.087)	0.115 (0.086)	0.120 (0.080)	0.112 (0.088)	0.118 (0.085)	0.118 (0.079)
30	-0.123 (0.020)	r	0.329 (0.102)	0.330 (0.102)	0.331 (0.103)	0.332 (0.103)	0.336 (0.103)	0.331 (0.101)	0.330 (0.104)	0.331 (0.102)	0.328 (0.104)
		k	0.208 (0.083)	0.206 (0.084)	0.207 (0.084)	0.207 (0.080)	0.212 (0.083)	0.213 (0.086)	0.208 (0.082)	0.214 (0.085)	0.211 (0.081)
50	-0.085 (0.014)	r	0.424 (0.098)	0.427 (0.100)	0.429 (0.099)	0.431 (0.099)	0.434 (0.098)	0.425 (0.099)	0.425 (0.097)	0.425 (0.098)	0.424 (0.098)
		k	0.274 (0.090)	0.277 (0.088)	0.279 (0.088)	0.280 (0.089)	0.283 (0.087)	0.276 (0.091)	0.278 (0.086)	0.277 (0.087)	0.278 (0.088)

() 内は標準偏差を表す

表 4-15　低識別力・中困難度の各指標(N=100)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.282 (0.049)	r	0.226 (0.106)	0.229 (0.109)	0.231 (0.112)	0.231 (0.112)	0.230 (0.111)	0.228 (0.108)	0.233 (0.105)	0.232 (0.107)	0.229 (0.105)
		k	0.135 (0.080)	0.137 (0.085)	0.138 (0.085)	0.136 (0.087)	0.136 (0.087)	0.136 (0.088)	0.131 (0.082)	0.130 (0.084)	0.141 (0.075)
30	-0.141 (0.025)	r	0.380 (0.103)	0.382 (0.104)	0.383 (0.106)	0.384 (0.106)	0.385 (0.107)	0.380 (0.102)	0.378 (0.103)	0.379 (0.102)	0.379 (0.104)
		k	0.236 (0.095)	0.237 (0.094)	0.236 (0.094)	0.235 (0.094)	0.233 (0.095)	0.233 (0.094)	0.235 (0.096)	0.237 (0.094)	0.231 (0.094)
50	-0.094 (0.016)	r	0.481 (0.090)	0.485 (0.090)	0.487 (0.091)	0.488 (0.091)	0.490 (0.092)	0.480 (0.089)	0.481 (0.089)	0.481 (0.089)	0.481 (0.089)
		k	0.304 (0.076)	0.308 (0.079)	0.313 (0.076)	0.313 (0.076)	0.310 (0.077)	0.305 (0.078)	0.307 (0.080)	0.306 (0.081)	0.311 (0.078)

() 内は標準偏差を表す

表 4-16　低識別力・高困難度の各指標(N=100)

項目数	Δ*D*		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.239 (0.043)	r	0.193 (0.111)	0.196 (0.109)	0.195 (0.111)	0.196 (0.112)	0.197 (0.114)	0.190 (0.115)	0.191 (0.110)	0.189 (0.113)	0.193 (0.110)
		k	0.120 (0.083)	0.117 (0.081)	0.119 (0.086)	0.122 (0.088)	0.116 (0.087)	0.115 (0.087)	0.119 (0.082)	0.114 (0.085)	0.120 (0.087)
30	-0.122 (0.021)	r	0.338 (0.109)	0.341 (0.109)	0.342 (0.110)	0.342 (0.109)	0.346 (0.112)	0.335 (0.108)	0.337 (0.109)	0.335 (0.108)	0.340 (0.110)
		k	0.202 (0.091)	0.205 (0.091)	0.208 (0.091)	0.210 (0.095)	0.217 (0.090)	0.209 (0.088)	0.204 (0.095)	0.209 (0.087)	0.208 (0.091)
50	-0.086 (0.015)	r	0.424 (0.101)	0.428 (0.103)	0.429 (0.103)	0.431 (0.103)	0.432 (0.103)	0.425 (0.100)	0.423 (0.100)	0.425 (0.099)	0.425 (0.100)
		k	0.270 (0.086)	0.277 (0.087)	0.277 (0.088)	0.276 (0.086)	0.273 (0.088)	0.270 (0.086)	0.270 (0.083)	0.271 (0.084)	0.274 (0.079)

() 内は標準偏差を表す

表 4-17　中識別力・困難度一様の各指標(N=100)

項目数	ΔD		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.178 (0.038)	r	0.444 (0.111)	0.448 (0.110)	0.451 (0.110)	0.452 (0.107)	0.456 (0.108)	0.446 (0.112)	0.446 (0.109)	0.444 (0.111)	0.446 (0.110)
		k	0.282 (0.094)	0.289 (0.097)	0.293 (0.091)	0.294 (0.089)	0.299 (0.092)	0.287 (0.096)	0.288 (0.093)	0.288 (0.099)	0.292 (0.099)
30	-0.072 (0.013)	r	0.678 (0.064)	0.680 (0.064)	0.680 (0.064)	0.681 (0.065)	0.685 (0.064)	0.677 (0.064)	0.679 (0.064)	0.678 (0.065)	0.678 (0.065)
		k	0.470 (0.071)	0.474 (0.067)	0.474 (0.068)	0.476 (0.068)	0.481 (0.072)	0.474 (0.068)	0.475 (0.071)	0.473 (0.070)	0.470 (0.071)
50	-0.047 (0.011)	r	0.767 (0.056)	0.770 (0.056)	0.771 (0.056)	0.772 (0.056)	0.775 (0.055)	0.768 (0.056)	0.768 (0.055)	0.768 (0.056)	0.768 (0.056)
		k	0.552 (0.063)	0.557 (0.060)	0.558 (0.064)	0.559 (0.065)	0.557 (0.065)	0.554 (0.061)	0.555 (0.061)	0.554 (0.060)	0.554 (0.064)

() 内は標準偏差を表す

表 4-18　中識別力・低困難度の各指標(N=100)

項目数	ΔD		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.172 (0.034)	r	0.453 (0.106)	0.455 (0.108)	0.456 (0.110)	0.458 (0.112)	0.461 (0.114)	0.454 (0.105)	0.453 (0.108)	0.453 (0.107)	0.451 (0.110)
		k	0.287 (0.088)	0.287 (0.090)	0.288 (0.089)	0.291 (0.090)	0.295 (0.084)	0.289 (0.095)	0.287 (0.093)	0.289 (0.095)	0.287 (0.095)
30	-0.073 (0.015)	r	0.670 (0.066)	0.674 (0.065)	0.676 (0.066)	0.677 (0.066)	0.682 (0.068)	0.671 (0.068)	0.669 (0.066)	0.670 (0.067)	0.670 (0.068)
		k	0.462 (0.075)	0.466 (0.077)	0.469 (0.075)	0.471 (0.075)	0.473 (0.080)	0.465 (0.077)	0.461 (0.075)	0.461 (0.078)	0.459 (0.075)
50	-0.048 (0.010)	r	0.752 (0.052)	0.754 (0.051)	0.754 (0.051)	0.754 (0.051)	0.757 (0.051)	0.750 (0.053)	0.751 (0.052)	0.751 (0.053)	0.749 (0.052)
		k	0.548 (0.073)	0.550 (0.071)	0.550 (0.069)	0.549 (0.071)	0.553 (0.069)	0.548 (0.073)	0.549 (0.070)	0.547 (0.073)	0.548 (0.069)

() 内は標準偏差を表す

表 4-19　中識別力・中困難度の各指標(N=100)

項目数	ΔD		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.188 (0.035)	r	0.517 (0.090)	0.522 (0.092)	0.524 (0.095)	0.526 (0.094)	0.530 (0.096)	0.516 (0.093)	0.518 (0.092)	0.517 (0.093)	0.521 (0.088)
		k	0.334 (0.087)	0.337 (0.079)	0.340 (0.078)	0.343 (0.074)	0.343 (0.078)	0.336 (0.082)	0.338 (0.083)	0.334 (0.089)	0.342 (0.076)
30	-0.077 (0.017)	r	0.729 (0.058)	0.732 (0.058)	0.734 (0.057)	0.735 (0.057)	0.737 (0.056)	0.730 (0.056)	0.729 (0.057)	0.730 (0.056)	0.727 (0.059)
		k	0.523 (0.076)	0.523 (0.077)	0.525 (0.077)	0.527 (0.076)	0.532 (0.077)	0.522 (0.078)	0.519 (0.075)	0.519 (0.075)	0.518 (0.075)
50	-0.049 (0.011)	r	0.809 (0.044)	0.811 (0.043)	0.811 (0.043)	0.812 (0.043)	0.815 (0.042)	0.809 (0.043)	0.809 (0.043)	0.809 (0.042)	0.809 (0.042)
		k	0.599 (0.068)	0.599 (0.065)	0.600 (0.063)	0.600 (0.061)	0.607 (0.067)	0.600 (0.063)	0.599 (0.064)	0.599 (0.064)	0.601 (0.064)

() 内は標準偏差を表す

表 4-20　中識別力・高困難度の各指標(N=100)

項目数	ΔD		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.176 (0.035)	r	0.449 (0.102)	0.451 (0.105)	0.454 (0.107)	0.457 (0.107)	0.461 (0.106)	0.449 (0.102)	0.447 (0.103)	0.447 (0.103)	0.449 (0.100)
		k	0.289 (0.084)	0.292 (0.090)	0.294 (0.086)	0.294 (0.083)	0.298 (0.084)	0.293 (0.086)	0.289 (0.087)	0.292 (0.086)	0.291 (0.088)
30	-0.073 (0.016)	r	0.660 (0.077)	0.662 (0.077)	0.664 (0.077)	0.666 (0.077)	0.669 (0.077)	0.659 (0.077)	0.660 (0.079)	0.660 (0.078)	0.661 (0.078)
		k	0.462 (0.084)	0.458 (0.084)	0.462 (0.082)	0.463 (0.080)	0.467 (0.078)	0.458 (0.081)	0.460 (0.080)	0.457 (0.081)	0.462 (0.078)
50	-0.047 (0.010)	r	0.758 (0.054)	0.761 (0.054)	0.762 (0.053)	0.763 (0.053)	0.765 (0.054)	0.758 (0.054)	0.759 (0.053)	0.759 (0.054)	0.759 (0.054)
		k	0.553 (0.072)	0.551 (0.072)	0.553 (0.069)	0.553 (0.069)	0.558 (0.068)	0.551 (0.071)	0.548 (0.070)	0.550 (0.070)	0.553 (0.068)

() 内は標準偏差を表す

表 4-21　高識別力・困難度一様の各指標(N=100)

項目数	ΔD		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.127 (0.030)	r	0.644 (0.082)	0.646 (0.083)	0.647 (0.083)	0.648 (0.081)	0.651 (0.079)	0.643 (0.079)	0.644 (0.081)	0.642 (0.081)	0.645 (0.079)
		k	0.433 (0.081)	0.439 (0.080)	0.440 (0.082)	0.443 (0.080)	0.452 (0.081)	0.437 (0.080)	0.440 (0.080)	0.438 (0.080)	0.437 (0.084)
30	-0.050 (0.012)	r	0.830 (0.036)	0.831 (0.037)	0.832 (0.037)	0.832 (0.037)	0.834 (0.037)	0.831 (0.037)	0.830 (0.036)	0.830 (0.036)	0.830 (0.036)
		k	0.615 (0.058)	0.619 (0.059)	0.618 (0.059)	0.620 (0.057)	0.622 (0.059)	0.616 (0.060)	0.619 (0.060)	0.616 (0.059)	0.616 (0.058)
50	-0.030 (0.008)	r	0.887 (0.027)	0.888 (0.027)	0.888 (0.027)	0.888 (0.027)	0.890 (0.027)	0.887 (0.027)	0.887 (0.028)	0.887 (0.028)	0.887 (0.027)
		k	0.703 (0.058)	0.706 (0.056)	0.706 (0.056)	0.705 (0.055)	0.702 (0.057)	0.703 (0.057)	0.700 (0.058)	0.701 (0.057)	0.699 (0.057)

() 内は標準偏差を表す

表 4-22　高識別力・低困難度の各指標(N=100)

項目数	ΔD		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.125 (0.025)	r	0.644 (0.082)	0.647 (0.084)	0.649 (0.084)	0.650 (0.084)	0.651 (0.085)	0.642 (0.081)	0.644 (0.079)	0.643 (0.080)	0.641 (0.082)
		k	0.440 (0.083)	0.443 (0.085)	0.444 (0.085)	0.450 (0.084)	0.444 (0.084)	0.431 (0.083)	0.437 (0.083)	0.431 (0.084)	0.435 (0.087)
30	-0.049 (0.011)	r	0.820 (0.041)	0.821 (0.040)	0.822 (0.039)	0.822 (0.039)	0.824 (0.039)	0.819 (0.042)	0.820 (0.041)	0.819 (0.042)	0.819 (0.041)
		k	0.613 (0.058)	0.613 (0.060)	0.613 (0.058)	0.615 (0.060)	0.620 (0.060)	0.615 (0.061)	0.613 (0.061)	0.612 (0.060)	0.614 (0.061)
50	-0.031 (0.009)	r	0.884 (0.027)	0.885 (0.028)	0.885 (0.027)	0.885 (0.027)	0.887 (0.027)	0.884 (0.028)	0.884 (0.027)	0.883 (0.027)	0.884 (0.027)
		k	0.681 (0.062)	0.684 (0.059)	0.687 (0.060)	0.687 (0.058)	0.688 (0.058)	0.681 (0.063)	0.683 (0.061)	0.681 (0.060)	0.684 (0.058)

() 内は標準偏差を表す

表 4-23　高識別力・中困難度の各指標(N=100)

項目数	ΔD		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.135 (0.028)	r	0.705 (0.064)	0.709 (0.064)	0.712 (0.063)	0.712 (0.063)	0.714 (0.062)	0.705 (0.064)	0.704 (0.064)	0.704 (0.063)	0.704 (0.064)
		k	0.496 (0.071)	0.502 (0.066)	0.506 (0.065)	0.503 (0.065)	0.503 (0.067)	0.498 (0.073)	0.501 (0.070)	0.503 (0.070)	0.494 (0.068)
30	-0.051 (0.014)	r	0.867 (0.032)	0.868 (0.032)	0.869 (0.032)	0.869 (0.031)	0.870 (0.030)	0.867 (0.032)	0.867 (0.031)	0.868 (0.031)	0.867 (0.032)
		k	0.674 (0.060)	0.677 (0.058)	0.677 (0.058)	0.678 (0.057)	0.678 (0.058)	0.672 (0.054)	0.675 (0.061)	0.674 (0.057)	0.677 (0.057)
50	-0.031 (0.009)	r	0.914 (0.021)	0.915 (0.020)	0.915 (0.020)	0.915 (0.020)	0.916 (0.020)	0.914 (0.020)	0.914 (0.021)	0.914 (0.020)	0.914 (0.021)
		k	0.739 (0.052)	0.740 (0.053)	0.739 (0.052)	0.738 (0.051)	0.740 (0.050)	0.738 (0.051)	0.741 (0.049)	0.740 (0.050)	0.736 (0.053)

() 内は標準偏差を表す

表 4-24　高識別力・高困難度の各指標(N=100)

項目数	ΔD		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	-0.123 (0.030)	r	0.635 (0.085)	0.637 (0.086)	0.639 (0.087)	0.640 (0.088)	0.642 (0.086)	0.633 (0.083)	0.635 (0.088)	0.633 (0.085)	0.636 (0.089)
		k	0.433 (0.086)	0.435 (0.086)	0.434 (0.086)	0.435 (0.090)	0.441 (0.089)	0.429 (0.087)	0.434 (0.087)	0.432 (0.085)	0.434 (0.091)
30	-0.048 (0.012)	r	0.822 (0.035)	0.824 (0.035)	0.825 (0.035)	0.825 (0.035)	0.826 (0.035)	0.823 (0.035)	0.823 (0.034)	0.824 (0.035)	0.823 (0.034)
		k	0.619 (0.063)	0.621 (0.065)	0.624 (0.061)	0.623 (0.061)	0.622 (0.066)	0.618 (0.064)	0.623 (0.059)	0.620 (0.061)	0.621 (0.059)
50	-0.031 (0.007)	r	0.880 (0.024)	0.881 (0.024)	0.881 (0.024)	0.882 (0.024)	0.883 (0.024)	0.880 (0.024)	0.881 (0.024)	0.880 (0.024)	0.880 (0.023)
		k	0.685 (0.055)	0.684 (0.055)	0.681 (0.056)	0.682 (0.054)	0.683 (0.057)	0.682 (0.055)	0.681 (0.054)	0.680 (0.054)	0.681 (0.057)

() 内は標準偏差を表す

4.4 実際のテストデータでの確認(研究 2-3)

目的

研究 2-3 では，シミュレーションで示された研究 2-2 の結果について，並び替え方法が実際のテスト場面でも有用であるかどうかの確認をする。具体的には，実際のテストデータについて，潜在特性推定値θを受検者の能力値の真値として代用し，θによる並び替え(群分け)と，研究 2-2 で示した素点を用いた並び替え(群分け)の一致率を検討する。

方法

（1）テストの実施

2019 年 4 月～6 月にかけて愛知県内の大学生 477 名(有効回答 453 名)にテストを行った。使用したテストは国語・数学・英語から成る 26 項目である。研究 1 と同一の受検者・テストを用いた。

（2）使用するテスト条件

全受検者を母集団として，10 名，40 名，100 名をサンプリングした。テストは，全項目を用いた 26 項目と，英語のみを用いた 12 項目の 2 パタンについて検討した。

（3）並び変える方法

研究 2-2 と同じ方法を用いた。

（4）シミュレーションの手続き

1. 全受検者，全項目について，2 PLM を用いて項目特性値ならびに潜在能力特性値を推定した。
2. 全受検者より N=10, 40, 100 についてサンプリングを行い，正答率ならびに I-T 相関を算出した。
3. サンプルを Type1～Type3.4 について順位付けならびに群分けを行い，D 指標を算出した。
4. 潜在特性推定値と比較するため，スピアマンの順位相関係数，重みづけカッパを算出した。
5. 2～4 について，全 26 項目ならびに英語のみ 12 項目のそれぞれについて 100 回行った。

（5）使用する指標

潜在特性推定値と提案手法での順序を比較検討するため，順位相関係数，重みづけカッパ係数について検討した。

結果

まず，使用したテストについて一次元性を確認するため，固有値についてλ_1/λ_2を確認し

た。26 項目のテストにおいて 2.359，12 項目のテストにおいて 1.731 と，どちらのテストも第一固有値が第二固有値に比べて大きく，一次元性が確認された。

それぞれのテストについて 2PLM を用いて推定したテスト項目の識別力，困難度の推定値ならびに I-T 相関と正答率は表 4-25,4-26 であった。

26 項目のテストについて，I-T 相関は M=0.259，SD=0.122，正答率は M=0.587，SD=0.201 であり，12 項目のテストについて，I-T 相関は M=0.198，SD=0.113，正答率は M=0.635，SD=0.261 であった。

これらのテストについて，N=10, 40, 100 をサンプリングして行ったシミュレーションの結果を表 4-27,4-28 に示した。項目特性推定値に基づき推定した潜在特性推定値を受検者の真値として代用し，順位相関係数ならびに重みづけカッパ係数を算出したものである。

26 項目のテストでは，受検者数によらず r=.921~.956，κ=.791~.834 と信頼性の基準を満たす値が得られた。並び替えのパタンでは，受検者の正答項目の識別力を用いて並び替える方法において最も高い重み付けカッパ係数の値となった。

表 4-25　26 項目のテスト項目特性推定値ならびに I-T 相関と正答率

	項目識別力	項目困難度	I-T相関	正答率
項目1	0.372	-1.185	0.141	0.605
項目2	0.258	-1.069	0.117	0.567
項目3	1.039	-1.023	0.342	0.704
項目4	1.019	-1.073	0.338	0.711
項目5	1.235	-0.637	0.383	0.645
項目6	0.896	0.125	0.295	0.475
項目7	1.053	-0.026	0.401	0.503
項目8	1.013	-0.239	0.344	0.547
項目9	0.127	7.687	0.035	0.274
項目10	1.119	-0.805	0.348	0.671
項目11	1.532	0.238	0.443	0.433
項目12	0.795	-0.128	0.297	0.521
項目13	0.955	-0.059	0.321	0.510
項目14	1.036	0.047	0.355	0.488
項目15	0.275	-0.973	0.127	0.565
項目16	0.281	4.701	0.083	0.214
項目17	0.510	1.214	0.191	0.358
項目18	1.190	-2.817	0.241	0.943
項目19	2.062	-2.288	0.259	0.960
項目20	0.903	-0.118	0.326	0.521
項目21	2.400	-1.680	0.359	0.914
項目22	-0.106	2.401	-0.048	0.563
項目23	1.113	-1.618	0.295	0.812
項目24	2.293	-1.691	0.369	0.912
項目25	0.482	-1.056	0.199	0.618
項目26	0.545	2.243	0.168	0.241

表 4-26　12 項目のテストの項目特性推定値ならびに I-T 相関と正答率

	項目識別力	項目困難度	I-T相関	正答率
項目15	0.318	-0.844	0.135	0.565
項目16	0.080	16.259	0.017	0.214
項目17	0.691	0.938	0.184	0.358
項目18	1.705	-2.286	0.306	0.943
項目19	1.978	-2.378	0.253	0.960
項目20	1.069	-0.097	0.247	0.521
項目21	2.052	-1.807	0.287	0.914
項目22	-0.127	1.993	-0.046	0.563
項目23	1.550	-1.325	0.283	0.812
項目24	2.231	-1.731	0.344	0.912
項目25	0.588	-0.884	0.210	0.618
項目26	0.687	1.839	0.151	0.241

表 4-27　実データ 26 項目の各指標

受検者数		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	r	0.925	0.925	0.923	0.925	0.932	0.923	0.924	0.921	0.930
		(0.060)	(0.061)	(0.062)	(0.062)	(0.062)	(0.062)	(0.062)	(0.064)	(0.060)
	k	0.806	0.806	0.800	0.809	0.809	0.791	0.800	0.797	0.794
		(0.229)	(0.229)	(0.232)	(0.229)	(0.238)	(0.226)	(0.232)	(0.227)	(0.251)
40	r	0.938	0.939	0.938	0.940	0.947	0.936	0.937	0.936	0.942
		(0.046)	(0.047)	(0.048)	(0.048)	(0.048)	(0.048)	(0.048)	(0.049)	(0.047)
	k	0.804	0.806	0.807	0.812	0.823	0.793	0.805	0.801	0.804
		(0.169)	(0.170)	(0.172)	(0.170)	(0.175)	(0.168)	(0.171)	(0.168)	(0.186)
100	r	0.945	0.946	0.947	0.948	0.956	0.944	0.945	0.944	0.949
		(0.039)	(0.040)	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)	(0.042)	(0.039)
	k	0.808	0.812	0.815	0.820	0.834	0.802	0.812	0.809	0.811
		(0.140)	(0.141)	(0.143)	(0.141)	(0.145)	(0.139)	(0.141)	(0.139)	(0.154)

() 内は標準偏差を表す

表 4-28　実データ 12 項目の各指標

受検者数		type1	type2.1	type2.2	type2.3	type2.4	type3.1	type3.2	type3.3	type3.4
10	r	0.786	0.787	0.785	0.788	0.815	0.785	0.808	0.804	0.829
		(0.124)	(0.127)	(0.132)	(0.131)	(0.129)	(0.129)	(0.115)	(0.113)	(0.091)
	k	0.581	0.588	0.597	0.600	0.619	0.566	0.600	0.578	0.606
		(0.239)	(0.242)	(0.247)	(0.250)	(0.252)	(0.242)	(0.250)	(0.231)	(0.215)
40	r	0.805	0.809	0.810	0.815	0.847	0.800	0.819	0.811	0.837
		(0.098)	(0.101)	(0.106)	(0.106)	(0.106)	(0.103)	(0.095)	(0.094)	(0.073)
	k	0.592	0.601	0.609	0.617	0.653	0.575	0.598	0.582	0.615
		(0.181)	(0.182)	(0.188)	(0.190)	(0.196)	(0.187)	(0.193)	(0.182)	(0.166)
100	r	0.816	0.820	0.824	0.830	0.865	0.816	0.827	0.825	0.843
		(0.083)	(0.086)	(0.091)	(0.091)	(0.092)	(0.091)	(0.082)	(0.083)	(0.063)
	k	0.596	0.604	0.614	0.622	0.667	0.587	0.598	0.593	0.617
		(0.152)	(0.153)	(0.157)	(0.159)	(0.166)	(0.161)	(0.165)	(0.158)	(0.140)

() 内は標準偏差を表す

12 項目のテストでは，受検者数によらず r=.785~865，κ=.578~.667 となった。並び替えのパタンでは，受検者の正答項目の識別力を用いて並び替える方法において最も高い重み付けカッパ係数の値となったが，基準となる 0.7 を超える条件はなかった。

考察

使用した 26 項目ならびに 12 項目のテストは，I-T 相関の平均値がそれぞれ 0.259，0.198 であることから，研究 2-2 で設定した識別力条件のうち，平均値が 0.25 となる中程度の条件に近かったと考えられる。また，正答率については，26 項目，12 項目のどちらも 0.214～0.960 の範囲であり，平均値ならびに標準偏差が 0.587(0.201)，0.635(0.261)であることから困難度が一様である条件に近かったと考えられる。

識別力が中程度かつ困難度が一様であるシミュレーションの結果(表 4-5，4-17)と比較すると，順位相関係数ならびに重み付けカッパ係数のどちらもが，シミュレーションよりも実データについて高い値を示した。ここでは，項目反応パタンを用いて 2PLM で推定した項目識別力ならびに項目困難度を算出した上で，推定された潜在能力推定値を受検者の能力値の真値として代用している。IRT モデルにおける推定値には誤差が含まれる(Chen & Wang, 2007)。推定は項目への反応パタンに基づいて行われており，よりデータに適合するように過剰推定したものとなっている可能性がある。そのため，真値と素点の並び替えを比較したシミュレーションの結果に比べて，能力推定値と素点の並び替えを比較するという実データを用いたシミュレーション結果の方が，並び順の一致率が高くなったと考えられる。

また，項目数に着目すると，26 項目のテストでは信頼性の基準を満たす k，r の値であったが，12 項目のテストではその基準は満たされなかった。このように実データを用いたシミュレーションからも，信頼性を上げるためには項目数の影響が大きいことが確認された。

合計得点が同じである受検者を並び替える方法として望ましいのは，項目の識別力を用いる方法(Type2.1～2.4)であった。中でも，受検者が正答した全ての項目の識別力について比較をし，識別力がより高い項目に正答した受検者の順に並び替えることが最も望ましい。しかし，テスト項目のうち，識別力の最も高い 1 項目についてのみの正誤について並べ替えても，ID 順に並べたときに比べて重み付けカッパ係数は大きくなった。

また，困難度を用いた並び替えでは，ID 順による並び替えよりも重み付けカッパ係数が低くなる条件もあった。具体的には，正答率が 0.5 に最も近い項目は 26 項目のテストにおいて項目 13，12 項目のテストにおいて項目 20 であり，そこで用いられる項目の I-T 相関係数はそれぞれ 0.321，0.247 と中程度の識別力であった。この 1 項目を用いて並び替えた結果が Type3.1 であるが，全ての条件において ID 順による並び替えよりも順位相関係数ならびに重み付けカッパ係数は低い値となった。つまり，識別力が十分でない項目の正誤情報を用いて並び替えると，いたずらに信頼性を下げることがあるといえる。そのため，困難度ではなく識別力を参照し，より識別力の高い項目の正誤情報を用いることが望ましいことが確認された。

4.5 考察

本研究では，テストの評価として *D* 指標を用いる際に望ましいテストの条件を確認すると共に，受検者の能力を適切に反映した群分けを行うための並び替えの手法について検討を行った。まず，カットオフポイント上の同じ得点の受検者について同じ群とする分け方を検討した上で，同点受検者を異なる群とする方法について検討し，実データを用いた確認を行った。その結果，望ましいとされるテストは，識別力が高く，かつ困難度が中程度に固まる項目数が 30 項目以上のときであることがわかった。また，項目数による影響が大きいことから，項目数が 50 項目以上であれば，更に真値に近い群分けを行うことができることがわかった。同点の受検者を分ける方法としては，受検者が正答した項目のうち，より識別力の高い項目に正答した順に並び替えた上で群分けを行うことが，より真値に近い分け方になることが明らかとなった。また，テスト項目の中で識別力が高かった 1〜3 項目の正誤について並び替えるだけでも，ランダムである ID 順よりは真値に近い並び順になった。

実際のクラスルームテスト場面におけるテスト時間は，50 分程度と限られた時間の中で実施される。科目によって項目を増やしやすい，増やしにくいといった違いはあるものの，真値に近い結果を得るために，闇雲にテストの項目数を増やすということはあまり現実的ではない。つまり，限られた項目数の中でよりテストの精度を上げることが望ましく，そのためには項目の識別力が高い項目を多く含む必要がある。項目の識別力に影響を及ぼす問題としては，赤根他(2006)で検討がされているものの，実施したテストを用いてどのような項目で識別力が高くなったかというものであり，識別力を高くするような作問方法については検討されていない。識別力の高い項目を効率的に作問するためにも，識別力へ影響すると考えられる要因を操作する等，より実証的な研究によって識別力が高い項目を作成する方法について検討していくことが今後の課題と考えられる。

第5章
総合的考察

5.1 テストの作成と評価に関する考察

本論文では，よいテストを作成するための方法として，テスト項目の作成時ならびにテスト実施後の評価にかかわる問題について検討を行った。特に，クラスルーム場面において教師がテストの作成・実施・分析を行う際に有用な知見を示すことを目的とした。

第 3 章では研究 1 として，テスト項目の作成に関する問題についての検討を行った。Haladyna & Rodriguez(2013)の示す多枝選択式項目作成ガイドラインの影響を実証的に検討するため，ガイドラインに準拠／非準拠となる項目を作成し，それぞれの正答率・識別力・コメント率の差について検討を行った。その結果，ガイドライン 3「各設問の内容は互いに独立であること」やガイドライン 18「『上記のいずれでもない』『上記すべてあてはまる』『分からない』などの選択枝を用いないこと」などの複数のガイドラインについて，コメント率が低いにもかかわらず，正答率や識別力に差が生じた。つまり，Flaw のある項目形式に受検者は違和感を覚えないにもかかわらず，測定したい能力とは別の特性によって正答率や識別力に影響を及ぼすということである。こうした項目では，受検者自身が気付き対処をすることが難しいため，特に注意が必要となる。

第 4 章では研究 2 として，テスト実施後の評価に関する問題についての検討を行った。具体的には，識別力を表す指標のうち，より簡便に扱うことが可能な *D* 指標について検討を行った。*D* 指標を扱う上で，受検者を得点に基づき 3 群に分ける必要がある。受検者の能力の真値をより反映するための群分け方法ならびに，それを可能とするテスト条件について検討した。その結果，各群の受検者を Kelly(1939)の割合となるよう，カットオフポイント上に居る複数の同点者についても異なる群として分けることが望ましいとされた。同点者を群分けする際には，識別力が高い項目の正誤を用いて並び替えた上での群分けを行うことが望ましいとされた。また，*D* 指標を算出するにあたり，望ましいテスト条件は，項目数が 30 項目以上かつ識別力が高い項目から構成され，困難度は中程度に固まるものであることが明らかとなった。

以上を踏まえると，クラスルームテストを行う上では識別力の高い項目を多く用いることが望ましい。また，項目を作成するにあたり，項目作成ガイドラインを参照することで，測定したい能力以外の要因の影響を避けることが求められる。特に，識別力について注意が必要である。たとえばガイドライン 3「各設問の内容は互いに独立であること」などの一部のガイドラインでは，準拠した方が項目の識別力が低くなるものが存在する。しかし，識別力の高い項目を作成するために，そうしたガイドラインに準拠しないことを推奨するものではない。ガイドラインに準拠しないことで識別力が高くなる要因は，測定したい能力とは異なる要因によるものである。項目が測定したい受検者の能力を十分に反映しない項目となることから，項目の形式を操作することで，識別力を不当に高めることは望ましくない。

なお，テストに記述式項目が含まれることで，結果的に全体の項目数が少なくなるという

ケースも考えられる。記述式項目を用いる理由によっては，多枝選択式項目を用いることが可能な場面も存在すると考えられる。たとえば，池田(1992)は，記述式の代わりに多枝選択式形式を用いるのであれば，評価の観点ごとに複数の項目を用いることで深い内容を引き出すことに繋がるという。しかし，記述式項目がテストに必要であるかどうかの検討は，テストの目的や測定したい能力に照らしながら行う必要がある。そのため，多枝選択式以外の項目形式についてもガイドラインの検討が必要だろう。

5.2 本研究の限界と展望

本研究は，クラスルームテストの作成・実施・分析について一定の知見を提供するものであると考えられる。一方で，今後検討されるべき課題も多く残されている。たとえば，項目 Flaw の存在を判断することの難しさや，科目の特徴を踏まえた検討，また，その他の項目形式についての検討はなされていない。

項目の Flaw の有無の判断の難しさ

本研究では，項目の Flaw の有無の判断方法については言及されていない。つまり，作成された項目について，項目分析の結果から Flaw の有無について判断できるものではない。

Flaw のない項目では，受検者の能力に基づいた正誤判断が行われている。しかし，作成した項目の識別力が高かったとしても，それだけをもって Flaw のなかった項目ということではない。一部の Flaw は項目の識別力を不当に高めることが確認されたことから，ガイドラインに則っていないことが識別力を高める要因にもなる。そのため，作成したテストについて，項目の正答率や識別力といった情報だけでその項目の良し悪しを判断することなく，ガイドラインに則った作題がなされているかどうかなどを確認し，総合的に判断する必要がある。

科目の特徴をふまえた議論の必要性

ここで検討された事項は，各科目の個別的な特徴を踏まえたものではない。ガイドラインに準拠することを重視しすぎることで，科目によっては不自然な項目形式となれば，受検者にとって違和感を抱かせることにも繋がる。そのため，実際にテストを作成する際には，ここで得られた知見をどこまで適用できるのかを十分に留意する必要がある。

たとえば，ひとつの長文を題材として，複数の項目が連なるような大問形式では，英語や国語といった科目においてよく見られる形式である。こうした大問形式の項目群では，問う内容によって項目間に依存性が生じる場合がある。ひとつの内容について関連づけられている複数の項目群では，その内容への事前知識が項目の正誤へ影響する可能性があり，このとき，一部の受検者に有利に働くなど，項目間の独立性が保たれなくなる(加藤・山田・川端, 2014)。また内容への事前知識が一定であっても，長文の中の単語の意味を答えさせる項目と，その単語を含む文章の意味を答えさせる項目でも，依存性が生じると考えられる(光永, 2017)。このように，大問形式の項目群は依存性が生じる項目を作成しやすい環境であり，ガイドライン 3「各設問の内容は独立であること」に反する可能性を孕んでいる。

しかしながら，大問形式それ自体が問題ということではない。多くの英語テストや国語テストでは，こうした大問形式を用いることの方が自然であり，また，複数の項目を作成するためにも必要であると考えられる。項目間の独立性を確保するために，ひとつの長文につきひとつの項目という作問を行うことの方が不自然であり，また，限られた試験時間内に多くの項目の解答を得るという観点からも現実的ではない。項目間に依存性を生じさせないた

めに，大問形式の項目群では，複数の項目間で解答に必要となる長文箇所が重ならないようなにするなどといった方法で対処をすることが現実的だと考えられる。

具体的な科目を踏まえ，それぞれのガイドラインがどのように影響を及ぼすかどうかについては，今後検討される必要がある。その際，教科科目としての名称による違いのみならず，科目ごとに求められる能力や主に用いられる項目形式の特徴を踏まえた上で検討することが望ましい。そのような検討を行うことで，特定の科目以外のテスト(e.g.大学における心理学関連の科目や資格試験，社内での昇進試験など)の作成にも知見を広げることが可能となる。

他の項目形式の検討の必要性

本研究では，多枝選択式項目作成ガイドラインについてのみ取り上げた。しかし，科目によっては記述式項目を含むなど，全ての項目を多枝選択式項目のみで作成するのは現実的ではないことがある。そのため，その他の項目形式においても同様に，項目作成ガイドラインの実証的な検討が待たれる。また，これらのガイドラインは国外で検討されてきたものがほとんどである。そのため，日本でのテスト文化を踏まえたガイドラインについても検討される必要があるだろう。

なお，文章を用いて回答を引き出す形式を用いる場面は学力テストの他にも心理学における尺度や社会調査などが存在する。アンケート調査項目へのガイドラインの発展性については，坪田・石井・荒井・安永・寺尾(2021)などで検討されている。

今後の展望として

このように，実際にテストを作成・実施した際には，その結果を踏まえ再度ガイドラインについて確認することや，作題の時点で個別の科目の特徴を踏まえることで，よいテストの作成に繋がる。また，今回は取り上げなかった多枝選択式形式以外の項目形式なども含め，ガイドラインの実証的な検討が待たれる。

しかし上述したように，本研究では科目の特徴を踏まえた検討は行っていない。科目が異なることによって，同じガイドラインであっても項目特性への影響の程度が異なる可能性がある。つまり，項目分析に求められるテスト条件を満たすために，科目によって参照するべきガイドラインが異なる可能性がある。更に，識別力の高い項目となる要因についても，科目によってその影響は異なる可能性がある。そのため今後の展望として，科目別に，項目作成ガイドラインの影響の多寡について科目の特徴を踏まえた整理をした上で，科目別のガイドラインが作成されることが求められる。加えて，異なる項目形式間でガイドラインを比較検討することも必要だろう。特に，複数の項目形式を用いることが多い科目では，ガイドラインによる影響のみならず，項目形式の違いによる影響も考慮する必要が生じる。

本研究ではガイドラインを検討するにあたり，正答率や識別力といった指標を取り上げた。しかし，項目を特徴づける指標は他にも存在しており(e.g.無答率,情報量)，他の指標を用いることで，ガイドラインの影響に関して本研究とは異なる視点から検討することが可

能だろう。

本論文では，クラスルームテストの作成・実施・評価に関する知見を提供することを目的として行った。ここで得られた知見を踏まえ，今後より詳細な議論が展開されることが求められる。

引用文献

赤根敦・伊藤圭・林篤裕・椎名久美子・大澤公一・柳井晴夫・田栗正章 (2006). 識別指数による総合試験問題の項目分析. 大学入試センター研究紀要, 35, 19-47.

荒井清佳 (2015). 多肢選択式問題を作成する上で大切なこと―問題作成の専門家に対する調査結果に基づいて―. 日本テスト学会誌,*11*,21-34.

Azeem, M. (2012). Development of math proficiency test using item response theory(IRT). PhD Thesis, University of Education, Lahore.

東洋 (2001). 子どもの能力と教育評価(第 2 版), 東京大学出版会

Breakall, J., Randles, C. Tasker, R (2019). Development and use of a multiple-choice item writing flaws evaluation instrument in the context of general chemistry. *Chemistry Education Research and Practice*, *20*, 369-382.

Brown, J. D. (1996). Testing in language problems. Prentice-Hall, Inc. [和田稔(訳) (1999). 言語テストの基礎知識―正しい問題作成・評価のために 大修館書店]

Chen, C., Wang, W. (2007). Effect of ignoring item interaction on item parameter estimation and detection of interacting items. *Applied psychological measurement*, *31*, 388-411.

Chittooran, M. M., & Miles, D. P. (2001, April). Test-taking skills for multiple-choice formats: Implications for school psychologists. *Paper presented at the annual conference of the National Association of School Psychologists*, Washington, DC.

Diederich, P. B. (1973). Short-cut statistics for teacher-made tests. Educational Testing Service, Princeton, N.J.

Downing, S. M. (2005). The effect of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Science Education*,*10*,133-143.

Ebel, R.L. & Frisbie, D.A. (1991) Essentials of Educational Measurement. 5th Edition, Prentice-Hall, Englewood Cliffs.

Haladyna, T. M., & Downing, S. M. (1989). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 1, 51-78.

Haladyna, T. M., & Downing, S. M. & Rodriguez, M. C. (2002). A revies of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15, 309-334.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

橋本重治 (1979). 新・教育評価法概説. 金子書房

橋本重治 (1981). 到達度評価の研究 ―その方法と技術―. 図書文化社

樋口太郎 (2021). 指導要録　田中耕治(編) よくわかる教育(第 3 版) ミネルヴァ書房 pp.158-159

池上真人 (2015). 多肢選択文法問題の設問形式に関する研究――択一式と複数選択式の解答プロセスに焦点をあてて――. 言語文化研究, *35*, 55-72.

池田央 (1973). テストII. 東京大学出版会.

池田央 (1982). テストと測定. 第一法規出版株式会社.

池田央 (1992). テストの科学–試験にかかわるすべての人に－ 日本文化科学社.

池田央 (1994). 現代テスト理論. 朝倉書店.

池田央 (2006). 学力テストの科学. 山森光陽･荘島宏二郎(編著) 学力いま, そしてこれから, ミネルヴァ書房

石井秀宗 (2007). 記述式問題における無回答に関連する要因の検討―群馬県児童生徒学力診断テスト小学 6 年生国語テストデータ分析の結果から― 日本テスト学会誌, 3, 60-70.

石井秀宗 (2020). 項目分析システム. 石井研究室 Retrieved from http://www.educa.nagoya-u.ac.jp/~ishii-h/test_system.html (2021 年 3 月 10 日)

印東太郎・牧田稔・肥田野直 (1950). 心理学的測定 統計調査テスト. 金子書房

Johnson, A. P. (1951). Notes on a suggested index of item validity: The U-L Index. *Journal of Educational Psychology*, 42, 499–504.

Jozefowicz, R.F., Koppen, B. M., Case, S., Galbraith, R., Swanson, D. & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine*, 77(2), 156-161.

梶谷真也・小林健太郎・鈴木史馬・中田勇人・盛本圭一 (2013). 成績順位の通知と学習意欲. 明星―明星大学明星教育センター研究紀要, 3, 101-110.

Kane, M, T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73.

加藤健太郎・山田剛史・川端一光 (2014). R による項目反応理論. オーム社

Kelly, T.L. (1939) The selection of upper and lower groups for the validation of test items, *Journal of Educational Psychology*, 30, 17-24.

木村直史・福島統・栗原敏・黒沢博身 (2000). A,K および X 形式からなる多肢選択問題における知識レベルの推定. 医学教育, *31*, 435-442.

北尾倫彦 (1960). ひらがな文と漢字まじり文の読みやすさの比較研究. 教育心理学研究, *7*, 195-199.

小泉理恵 (2017). テストに必要な要素：妥当性, 信頼性, 実用性. 小泉理恵・印南洋・深澤真(編) 実例でわかる英語テスト作成ガイド. 大修館書店. pp.55-59

Lailamsyah, A. F. & Apriyanti, F. (2020). An item analysis of English summative test for the tenth grade students of sma muhammadiyah 3 Jakarta in the 2013/2014 academic year. Ed-Humanistics, 5, 610-615.

Lord, F. M. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics*, 7(3), 165-174.

Lord, F. M., & Novick. M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison Wesley.

松井仁 (1991). 項目反応理論の応用 II ロールシャッハテストの運動反応の数量化 芝祐

順(編) 項目反応理論－基礎と応用－東京大学出版会. pp187-194.

Martínez, R.J, Moreno, R., Martín, I., Trigo, M.E. (2009). Evaluation of five guidelines for option development in multiple-choice item-writing. *Psicothema, 21*, 326-330.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education. (池田央・柳井晴夫・藤田恵璽・繁枡算男(訳編) (1991). 『教育測定学(上巻)』 C.S.L.学習評価研究所. みくに出版)

光永悠彦 (2017). テストは何を測るのか 項目反応理論の考え方. ナカニシヤ出版

宮本友弘・倉元直樹 (2017). 国立大学における個別学力試験の解答形式の分類. 日本テスト学会誌,13,69-84.

日本テスト学会(編) (2007). テストスタンダード 日本のテストの将来に向けて 金子書房.

日本テスト学会(編) (2010). 見直そう，テストを支える基本の技術と教育. 金子書房

野口博之・大隅敦子 (2014). テスティングの基礎理論. 研究社.

大津明夫 (2011). テスト理論 松原望・美添泰人・岩崎学・金明哲・竹村和久・林文・山岡和枝(編), 統計応用の百科事典, 丸善出版, pp.420-421.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement, 24*, 3-14.

Rosenthal, R., Jacobson, L. (1968). Pygmalion in the classroom. The Urban Review, *3*, 16-20.

Rush, B. R., Rankin, D. C., White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education, 16*, 250-259.

斎藤泰一・有田清三 (1981). 多肢選択問題の正答率に及ぼす不十分な知識の影響. 医学教育, *12*, 356-360.

篠塚勝正・窪田三喜夫 (2012). 日本語文字形態(漢字，ひらがな，カタカナ)による認知言語処理の差異. 成城文芸，*221*，98-84.

Shizuka, T., Takeuchi, O., Yashima, T. Yoshizawa, K. (2006). A comparison of three- and four-option English tests for university entrance selection purpose in Japan. *Language Testing, 23*,35-37.

荘島宏二郎 (2010), 古典的テスト理論―科学的対象としてのテスト得点― 植野真臣・荘島宏二郎(著) 学習評価の新潮流, 朝倉書店, pp37-55.

Sireci, S. G., Yang, Y., Harter, J., Ehrlich, E. J. (2006). Evaluation guidelines for test adaptations A methodological analysis of translation quality. *Journal of Cross-Cultual Psychology, 37*, 557-567.

田中耕治 (2002). 到達度評価の理論と授業の改善 遠藤光男・天野正輝(編) 到達度評価の理論と実践. 昭和堂. pp61-75

田中耕治 (2008). 教育評価. 岩波書店

Tarrant, M., Knierim, A. Hayes, S.K. Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today, 26*, 662-671.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student

achievement in high-stakes nursing assessments. *Medical Education*, 42, 198-206.
寺尾尚大・安永和央・石井秀宗・野口裕之 (2015). 能力別にみた錯乱枝の効果に関する検討. 日本テスト学会誌, 11, 1-20.
遠山光則・中村弘明 (2013). 多肢選択式問題を考える：同一問題を A 形式、X2 形式、XX 形式で出題したときの正答率・識別係数・選択肢別解答率の比較. 東京歯科大学教養系研究紀要, *28*, 29-42.
坪田彩乃・石井秀宗 (2020). 多枝選択式問題作成ガイドラインの実証的検討. 日本テスト学会誌, *16*, 1-12.
坪田彩乃・石井秀宗 (印刷中). D 指標を用いた項目分析のためテスト条件の検討 ――より真値に近い状態を目指して. Quality Education
坪田彩乃・石井秀宗・荒井清佳・安永和央・寺尾尚大 (2021). アンケート調査項目作成ガイドラインの開発（1）― 日本での普及に向けた整理 ― 日本テスト学会第 19 回大会, オンライン
若林身歌 (2021). 目標に準拠した評価　田中耕治(編) よくわかる教育評価(第 3 版) ミネルヴァ書房 pp.24-25
若林俊輔・根岸雅史 (1993). 無責任なテストが「落ちこぼれ」を作る 正しい問題作成への英語授業学的アプローチ. 大修館書店
Wiyasa, P.I., Laksana, I.K.D, & Indrawati, N.L.K.M. (2019), Evaluating Quality of Teacher-Developed English Test in Vocational High School: Content Validity and Item Analysis. Education Quarterly Reviews, 2, 344-356.
八木英二 (2006). 到達度評価 辰野千壽・石田恒好・北尾倫彦(監) 教育評価事典 図書文化社 p.45
矢口新 (1957). 基礎学力の診断. 法政大学出版局
安永和央・斎藤信・石井秀宗 (2010). 構造的性質を操作した国語テストにおける回答の検討―中学生を対象にしたテストの実証研究― 日本テスト学会誌, 8, 117-132.
安永和央・石井秀宗 (2012).テストにおける設問の問い方が回答傾向に及ぼす影響. 教育心理学研究, *60*, 296-309.
Yu-mien, S. (2010). An item analysis of an English achievement test taken by EFL college students in Taiwan. 明道學術論壇, 6(3), 59-82.

Appendix

実際に使用したテスト冊子

テストへのご協力のお願い

このテストは，より良いテスト問題を作る研究に役立てるためのものです。解答は統計的に処理されるため，個人を特定する目的では使用いたしません。また，解答をしなかったり，途中で解答をやめたりしても一切の不利益はありません。いただいた解答は，個人が特定されない形でUSB メモリに入力後，パスワードをかけて厳重に保管し，大学の定める期間終了後に責任をもって処分をします。研究の結果は，個人情報が特定されないように処理をした上で，学会や学術雑誌等で発表されることがあります。

このテストに**ご協力いただける方は，先に進んでください。**

解答用紙の提出をもって，協力への同意の確認とさせていただきます。

同意を撤回する場合，結果のフィードバックが必要な場合，ご質問がある場合には以下へお問い合わせください。

【研究実施者】

石井秀宗(名古屋大学大学院教育発達科学研究科　教授)

ishii.hidetoki@b.mbox.nagoya-u.ac.jp

坪田彩乃(名古屋大学大学院教育発達科学研究科　博士後期課程 3 年)

ayano.melon@gmail.com

【国語・数学・英語】
(30 分)

《 注 意 事 項 》

1．試験開始の合図があるまでこの問題冊子を開いてはいけません。
2．問題冊子は 3〜10 ページに印刷されており，解答用紙 1 枚が挟まれています。
3．試験開始の合図があったら，解答用紙に学籍番号及び氏名を記入してください。
4．試験問題は全て選択問題です。当てはまる数字を解答用紙の解答欄に記入してください。
5．問題冊子の余白には書き込みができますが，**解答用紙の余白には書き込みできません。**
6．試験中に**質問はできません。**試験問題等について気が付いたことや不明な点は，**できるだけたくさん解答用紙の裏面に記入してください。**
7．問題冊子は持ち帰ってください。

【国語】

1. 近代文学史の啓蒙期に，当時一流の洋学者が集まって発行したものは何か。
   1. 我楽多文庫
   2. 明六雑誌
   3. しがらみ草紙
   4. 文学界

2. 『山月記』について正しいものを選べ。
   1. 全文に渡り漢詩で記述されている
   2. 中国の古典文学である
   3. 虎になった人物は李徴である
   4. 主人公はもともと上級官吏であった

3. 次の文の書き下し文として正しいものを選べ。
   死馬且買之。況生者乎。　十八史略・春秋戦国
   （死んだ馬でさえもなおかつ買うのだ。まして生きた馬ならばなおさら買うのだ。）
   1. 死馬は之を買ふ。況んや生者をや。
   2. 死馬は且つ之を買ふ。況くんぞ生ける者をや。
   3. 死馬すら且つ之を買ふ。況んや生ける者をや。
   4. 死馬すら且つなほ之を買ふ。況んや生ける者をや。

6. 「大江山　いく野の道の　遠ければ　まだふみもみず　天の橋立」　小式部内侍
下線部で用いられている修辞法は掛詞である。この修辞法について**正しくないもの**を選べ。
1. 和歌で用いられる修辞法である。
2. 発音が同じ言葉に二つ以上の意味を持たせるものである。
3. 用いることで内容を豊かに表現することができる。
4. 仮名の登場以前から多く使われていた。

7. 和歌の修辞法について正しいものを選べ。
1. 句切れとは，第五句以外の句の終わりに意味上の切れ目があるものである。
2. 枕詞とは，主に五音以外で特定の語を導き出すことを目的とした決まった言葉である。
3. 縁語とは，ある言葉と意味の上で関連しない言葉を連想的に用いる技法のことである。
4. 本歌取りとは，有名でない古歌から一～二句を取って新しい歌を詠む技法である。

8. 男性作者が女性になりきり執筆した，仮名文字を用いた日記形式の文学作品は何か。
1. 紫式部日記
2. 蜻蛉日記
3. 土佐日記
4. 中学聖日記

4. 次の文の書き下し文として正しいものを選べ。

天帝使我長百獣。　　戦国策・祖策

（天帝は私を多くの獣のかしらとした。）

1. 天帝百獣をして我に長たらしむ
2. 天帝我をして百獣に長たらしむ
3. 天帝長をして我に百獣たらしむ

5. 次の文を読んで，後の問いに答えよ。

下野の国に男女すみわたりけり。としごろすみけるほどに，男，妻まうけて心かはりはてて，この家にありける物どもを，今の妻のがりかきはらひもて運び行く。心憂しとおもへど，なほまかせてみけり。ちりばかりのものも残さずみな持て往ぬ。

『大和物語』百五十七段

下線部の訳として適当なものを，次の中から一つ選べ。

1. おかしいことが起こっている
2. つらい
3. 悲しい思いを私は今している
4. いまいましい感情である

【数学】

９．0.06^{28} は小数第何位に初めて 0 でない数字が現れるか。
ただし，$\log_{10}2=0.3010$, $\log_{10}3=0.4771$ とする。

１．33
２．34
３．35
４．36

１０．(1)　3 点(0,3)，(-2,17)，(1,5)を通る二次関数 $f(x)$を求めよ。

１．$f(x)=x^2+x+3$
２．$f(x)=2x^2+3$
３．$f(x)=3x^2-x+3$
４．$f(x)=4x^2-2x+3$

(2)　(1)で求めた二次関数の範囲が$-3 \leqq x \leqq 4$ のとき，y の最小値として正しいものを選べ。

１．35／12
２．28／9
３．47
４．33

１１．　はんけい 7 の　えん い と　はんけい 5 の　えん ろ が　ことなる　にてんで
まじわり　ふたつの　えんの　ちゅうしんかんのきょりを は とする
このとき　は の　とりうるあたいの　はんいを　もとめよ。

１．は しょうなり 2　　12 しょうなり は
２．2 しょうなり は しょうなり 12
３．は しょうなりいこおる 2　　12 しょうなりいこおる は
４．2 しょうなりいこおる は しょうなりいこおる 12

１２．　□に当てはまるものを選べ。

x,y を実数としたとき, $x>1$ かつ $y>2$ は, $xy>2$ かつ $x+y>3$ であるための□。

１．必要条件である

２．十分条件である

３．必要十分条件である

４．必要条件でも十分条件でもない

１３．　$x^2+5x-7=0$ のとき，解はどのようになるか以下の中から選べ。

１．　異なる 2 つの実数解

２．　重解

３．　異なる 2 つの虚数解

４．　1 つの実数解と 1 つの虚数解

【英語】

１４． 下線部の発音がほかの三つと異なるものを，１～４の中から一つ選べ。

１．m<u>ou</u>se

２．l<u>ou</u>nge

３．c<u>ou</u>rage

４．p<u>ou</u>nd

１５． “diploma”の単語の意味を選べ。

１．交渉術

２．卒業証書

３．古門書学

４．外交官

１６． “issue”の意味として**<u>正しくないもの</u>**を選べ。

１．発行する

２．問題

３．(雑誌などの)号

４．学説

１７． 以下の説明が示すものを選べ。

a shop where you can buy food, alcohol, magazines etc, that is often open 24 hours each day

１．Park

２．Convenience store

３．City hall

４．Lawson

１８． 以下の説明が示すものを選べ。

a large park with many special machines that you can ride on, such as roller coasters and merry-go-rounds

１．Museum
２．Amusement park
３．Nagoya University
４．Aquarium

１９． 以下の文章の下線部と同じ意味を表すものを選べ。

Don’t <u>lose your temper</u>. It won’t help you.

１．catch cold
２．be sad
３．get angry
４．feel disappointed

２０． (　)に当てはまるものを選べ。

You should not depend (　　) your parents.

１．only
２．son
３．on
４．won

２１． (　)に当てはまるものを選べ。

(　　) is the best actor in the world.

１．Leonardo DiCaprio
２．Mickey Mouse
３．Harrison Ford
４．Simon Baker

２２．（　）に当てはまるものを選べ。

Alan Menken (　　) is a composer won an Academy Award for the Original Music Score of “Beauty and the Beast”.

１．that
２．which
３．what
４．who

２３．(　) に当てはまるものを選べ。

My mother allowed　(　　) to go to the Magic Kingdom with my friends.

１．I
２．my
３．me
４．mine

２４．以下の会話を完成させよ。

A : I persuaded him to buy a new wallet.
B : Wow! How did you do?
A : (　　)
B : That`s awful!

１．I broke his wallet.
２．It was 15,000 yen.
３．He used all of his money.
４．His wallet was terrible.

２５．以下の会話を成立させよ。

A : Please call me Taxi.
B : (　　)

１．You should go strait.
２．Don’t you have a cell-phone?
３．OK, Taxi. That’s right?
４．OK, I will call.

解　答　用　紙

学籍番号 _____

氏名 _____

【 国 語 】

		3	2	1
8	7	6	5	4

【 数 学 】

9	１０(1)	１０(2)	１１
１２	１３		

【 英 語 】

１４	１５	１６	１７
１８	１９	２０	２１
２２	２３	２４	２５

試験問題等で気が付いたことを裏面に記入してください。

試験問題等について，気が付いたことや不明な点をできるだけたくさん記入してください。試験問題についての記述は，問題番号も記すようにしてください。

テストへのご協力のお願い

このテストは，より良いテスト問題を作る研究に役立てるためのものです。解答は統計的に処理されるため，個人を特定する目的では使用いたしません。また，解答をしなかったり，途中で解答をやめたりしても一切の不利益はありません。いただいた解答は，個人が特定されない形でUSB メモリに入力後，パスワードをかけて厳重に保管し，大学の定める期間終了後に責任をもって処分をします。研究の結果は，個人情報が特定されないように処理をした上で，学会や学術雑誌等で発表されることがあります。

このテストに**ご協力いただける方は，先に進んでください。**

解答用紙の提出をもって，協力への同意の確認とさせていただきます。

同意を撤回する場合，結果のフィードバックが必要な場合，ご質問がある場合には以下へお問い合わせください。

【研究実施者】

石井秀宗(名古屋大学大学院教育発達科学研究科　教授)

ishii.hidetoki@b.mbox.nagoya-u.ac.jp

坪田彩乃(名古屋大学大学院教育発達科学研究科　博士後期課程 3 年)

ayano.melon@gmail.com

【国語・数学・英語】
(30 分)

《 注 意 事 項 》

1．試験開始の合図があるまでこの問題冊子を開いてはいけません。
2．問題冊子は 3〜10 ページに印刷されており，解答用紙 1 枚が挟まれています。
3．試験開始の合図があったら，解答用紙に学籍番号及び氏名を記入してください。
4．試験問題は全て選択問題です。当てはまる数字を解答用紙の解答欄に記入してください。
5．問題冊子の余白には書き込みができますが，**解答用紙の余白には書き込みできません。**
6．試験中に**質問はできません。**試験問題等について気が付いたことや不明な点は，**できるだけたくさん解答用紙の裏面に記入してください。**
7．問題冊子は持ち帰ってください。

【国語】

1. 近代文学史の啓蒙期にあたる明治六年に，当時一流の洋学者が集まって発行したものは何か。
   1. 我楽多文庫
   2. 明六雑誌
   3. しがらみ草紙
   4. 文学界

2. 『山月記』について正しいものを選べ。
   1. 作者は芥川龍之介である
   2. 中国の古典文学である
   3. 虎になった人物は李徴である
   4. 虎になった人物の親友は李徴である

3. 次の文の書き下し文として正しいものを選べ。

庸夫且知其不可，況賢人乎。　　虞氏春秋

（凡夫でさえもなおかつそれが出来ないと知っているのだ。まして賢人ならばなおさら知っているのだ。）

   1. 庸夫は其の可ならざるを知る，況んや賢い人をや。
   2. 庸夫は且つ其の可ならざるを知る，況んぞ賢人をや。
   3. 庸夫すら且つ其の可ならざるを知る，況んや賢人をや。
   4. 庸夫すら且つなほ其の可ならざるを知る，況んや賢人をや。

6. 「大江山　いく野の道の　遠ければ　まだふみもみず　天の橋立」　小式部内侍
下線部で用いられている修辞法は掛詞である。この修辞法について正しいものを選べ。
1. 和歌で常に用いられる修辞法である。
2. 発音が同じ言葉に二つだけの意味を持たせるものである。
3. 必ず景物と心情の言葉を掛け合わせる。
4. 仮名の登場以降に多く使われるようになった。

7. 和歌の修辞法について**誤っているもの**を選べ。
1. 句切れとは，第五句の終わりに意味上の切れ目があるものである。
2. 枕詞とは，主に五音で特定の語を導き出すことを目的とした決まった言葉である。
3. 縁語とは，ある言葉と意味の上で関連する言葉を連想的に用いる技法のことである。
4. 本歌取りとは，有名な古歌から一～二句を取って新しい歌を詠む技法である。

8. 男性作者が女性になりきり執筆した，仮名文字を用いた日記形式の文学作品は何か。
1. 紫式部日記
2. 蜻蛉日記
3. 土佐日記
4. 更級日記

4. 次の文の書き下し文として正しいものを選べ。

天帝使我長百獣。　　戦国策・祖策

（天帝は私を多くの獣のかしらとした。）

1. 天帝百獣をして我に長たらしむ
2. 天帝我をして百獣に長たらしむ
3. 天帝長をして我に百獣たらしむ
4. 天帝長をして百獣に我たらしむ

5. 次の文を読んで，後の問いに答えよ。

下野の国に男女すみわたりけり。としごろすみけるほどに、男、妻まうけて心かはりはてて、この家にありける物どもを、今の妻のがりかきはらひもて運び行く。心憂しとおもへど、なほまかせてみけり。ちりばかりのものも残さずみな持て往ぬ。

『大和物語』百五十七段

下線部の訳として適当なものを，次の中から一つ選べ。

1. おかしい
2. つらい
3. 悲しい
4. いまいましい

【数学】

９．0.06^{28} は小数第何位に初めて 0 でない数字が現れるか。

１．33

２．34

３．35

４．36

１０．(1)　3 点(0,3)，(-2,17)，(1,5)を通る二次関数 $f(x)$を求めよ。

１．$f(x)=x^2+x+3$

２．$f(x)=2x^2+3$

３．$f(x)=3x^2-x+3$

４．$f(x)=4x^2-2x+3$

(2)　二次関数 $y=2x^2-5x+5$ の範囲が$-3 \leqq x \leqq 4$ のとき，y の最小値として正しいものを選べ。

１．35／16

２．15／8

３．17

４．38

１１．　半径 7 の円 P と半径 5 の円 Q が異なる 2 点で交わり，二つの円の中心間の距離を d とする。このとき，d の取りうる値の範囲を求めよ。

１．$d < 2$，$12 < d$

２．$2 < d < 12$

３．$d \leqq 2$，$12 \leqq d$

４．$2 \leqq d \leqq 12$

１２．　世の中には様々な条件が存在する。それは，数学の世界でも同様である。ここで，x,y を実数だとする。そこで，x と y についての条件を考えてみる。x と y の数値との関係性を考えたとき，$x>1$ かつ $y>2$ としよう。このとき，$x>1$ かつ $y>2$ という条件は $xy>2$ かつ $x+y>3$ であるための必要条件であるか，十分条件であるか，必要十分条件であるか，必要条件でも十分条件でもないかのいずれかである。当てはまるものを以下の選択枝の中から選べ。

１．$x>1$ かつ $y>2$ という条件は $xy>2$ かつ $x+y>3$ であるための必要条件である

２．$x>1$ かつ $y>2$ という条件は $xy>2$ かつ $x+y>3$ であるための十分条件である

３．$x>1$ かつ $y>2$ という条件は $xy>2$ かつ $x+y>3$ であるための必要十分条件である

４．$x>1$ かつ $y>2$ という条件は $xy>2$ かつ $x+y>3$ であるための必要条件でも十分条件でもない

１３．　$x^2+5x-7=0$

１．　異なる 2 つの実数解

２．　重解

３．　異なる 2 つの虚数解

４．　1 つの実数解と 1 つの虚数解

【英語】

１４． 下線部の発音がほかの三つと同じではないものを，１～４の中から一つ選べ。

１．m<u>ou</u>se

２．l<u>ou</u>nge

３．c<u>ou</u>rage

４．p<u>ou</u>nd

１５． “pneumonoultramicroscopicsilicovolcanoconiosis”の単語の意味を選べ。

１．肺炎

２．塵肺症

３．一過性脳虚血発作

４．副甲状腺機能亢進症

１６． “issue”の意味として正しいものを選べ。

１．発行する

２．問題

３．(雑誌などの)号

４．上記すべてあてはまる

１７． 以下の説明が示すものを選べ。

a shop where you can buy food, alcohol, magazines etc, that is often open 24 hours each day

１．Park

２．Convenience store

３．City hall

４．Fire Station

１８． 以下の説明が示すものを選べ。

a large park with many special machines that you can ride on, such as roller coasters and merry-go-rounds

１．Museum
２．Amusement park
３．Movie theater
４．Aquarium

１９． 以下の文章の下線部について，入れ替えができるものを選べ。

Don’t lose your temper. It won’t help you.

１．catch cold
２．be sad
３．get angry
４．feel disappointed

２０． (　)に当てはまるものを選べ。

You should not depend (　　) your parents.

１．in
２．at
３．on
４．to

２１． (　)に当てはまるものを選べ。

(　　) is the host of Tokyo Disney Resort.

１．Leonardo DiCaprio
２．Mickey Mouse
３．Harrison Ford
４．Simon Baker

２２．（　）に当てはまるものを選べ。

Alan Menken (　　) is a composer won an Academy Award for the Original Music Score of “Beauty and the Beast”.

１．whom
２．which
３．what
４．who

２３．(　) に当てはまるものを選べ。

My mother allowed　(　　) to go to the Magic Kingdom with my friends.

１．mine
２．me
３．my
４．I

２４．以下の会話を完成させよ。

A : I persuaded him to buy a new wallet.
B : Wow! How did you do?
A : (　　)
B : That`s awful!

１．I broke his wallet.
２．It was 15 yen.
３．He used all of his money.
４．His wallet was terrible.

２５．以下の会話を成立させよ。

A : Please call me Ann.
B : (　　)

１．You should go strait.
２．Don’t you have a cell-phone?
３．OK, Ann. That’s right?
４．OK, I will call.

解　答　用　紙

学籍番号 _____

氏名 _____

【 国 語 】

8	7	6	5	4

3	2	1

【 数 学 】

9	１０(1)	１０(2)	１１

１２	１３

【 英 語 】

１４	１５	１６	１７

１８	１９	２０	２１

２２	２３	２４	２５

試験問題等で気が付いたことを裏面に記入してください。

試験問題等について，気が付いたことや不明な点をできるだけたくさん記入してください。試験問題についての記述は，問題番号も記すようにしてください。

テストへのご協力のお願い

このテストは，より良いテスト問題を作る研究に役立てるためのものです。解答は統計的に処理されるため，個人を特定する目的では使用いたしません。また，解答をしなかったり，途中で解答をやめたりしても一切の不利益はありません。いただいた解答は，個人が特定されない形でUSB メモリに入力後，パスワードをかけて厳重に保管し，大学の定める期間終了後に責任をもって処分をします。研究の結果は，個人情報が特定されないように処理をした上で，学会や学術雑誌等で発表されることがあります。

このテストに**ご協力いただける方は，先に進んでください。**

解答用紙の提出をもって，協力への同意の確認とさせていただきます。

同意を撤回する場合，結果のフィードバックが必要な場合，ご質問がある場合には以下へお問い合わせください。

【研究実施者】

石井秀宗(名古屋大学大学院教育発達科学研究科　教授)

ishii.hidetoki@b.mbox.nagoya-u.ac.jp

坪田彩乃(名古屋大学大学院教育発達科学研究科　博士後期課程 3 年)

ayano.melon@gmail.com

【国語・数学・英語】
(30 分)

《 注 意 事 項 》

1．試験開始の合図があるまでこの問題冊子を開いてはいけません。
2．問題冊子は 3〜10 ページに印刷されており，解答用紙 1 枚が挟まれています。
3．試験開始の合図があったら，解答用紙に学籍番号及び氏名を記入してください。
4．試験問題は全て選択問題です。当てはまる数字を解答用紙の解答欄に記入してください。
5．問題冊子の余白には書き込みができますが，**解答用紙の余白には書き込みできません。**
6．試験中に**質問はできません。**試験問題等について気が付いたことや不明な点は，**できるだけたくさん解答用紙の裏面に記入してください。**
7．問題冊子は持ち帰ってください。

【国語】

1. 近代文学史の啓蒙期に，当時一流の洋学者が集まって発行したものは何か。
   1. 我楽多文庫
   2. 明六雑誌
   3. しがらみ草紙
   4. 文学界

2. 『山月記』について正しいものを選べ。
   1. 全文に渡り漢詩で記述されている
   2. 中国の古典文学である
   3. 虎になった人物は李徴である
   4. 主人公はもともと上級官吏であった

3. 次の文の書き下し文として正しいものを選べ。

   庸夫且知其不可，況賢人乎。　　虞氏春秋

（凡夫でさえもなおかつそれが出来ないと知っているのだ。まして賢人ならばなおさら知っているのだ。）

   1. 庸夫は其の可ならざるを知る，況んや賢い人をや。
   2. 庸夫は且つ其の可ならざるを知る，況んぞ賢人をや。
   3. 庸夫すら且つ其の可ならざるを知る，況んや賢人をや。
   4. 庸夫すら且つなほ其の可ならざるを知る，況んや賢人をや。

6. 「大江山　いく野の道の　遠ければ　まだふみもみず　天の橋立」　小式部内侍
下線部で用いられている修辞法は掛詞である。この修辞法について**正しくないもの**を選べ。
1. 和歌で用いられる修辞法である。
2. 発音が同じ言葉に二つ以上の意味を持たせるものである。
3. 用いることで内容を豊かに表現することができる。
4. 仮名の登場以前から多く使われていた。

7. 和歌の修辞法について**誤っているもの**を選べ。
1. 句切れとは，第五句の終わりに意味上の切れ目があるものである。
2. 枕詞とは，主に五音で特定の語を導き出すことを目的とした決まった言葉である。
3. 縁語とは，ある言葉と意味の上で関連する言葉を連想的に用いる技法のことである。
4. 本歌取りとは，有名な古歌から一～二句を取って新しい歌を詠む技法である。

8. 男性作者が女性になりきり執筆した，仮名文字を用いた日記形式の文学作品は何か。
1. 紫式部日記
2. 蜻蛉日記
3. 土佐日記
4. 更級日記

4. 次の文の書き下し文として正しいものを選べ。

天帝使我長百獣。　　戦国策・祖策

（天帝は私を多くの獣のかしらとした。）

1. 天帝百獣をして我に長たらしむ
2. 天帝我をして百獣に長たらしむ
3. 天帝長をして我に百獣たらしむ

5. 次の文を読んで，後の問いに答えよ。

下野の国に男女すみわたりけり。としごろすみけるほどに、男、妻まうけて心かはりはてて、この家にありける物どもを、今の妻のがりかきはらひもて運び行く。心憂しとおもへど、なほまかせてみけり。ちりばかりのものも残さずみな持て往ぬ。

『大和物語』百五十七段

下線部の訳として適当なものを，次の中から一つ選べ。

1. おかしい
2. つらい
3. 悲しい
4. いまいましい

【数学】

９．0.06^{28} は小数第何位に初めて 0 でない数字が現れるか。
ただし，$\log_{10}2=0.3010$, $\log_{10}3=0.4771$ とする。

１．33
２．34
３．35
４．36

１０．(1)　3 点(0,3)，(-2,17)，(1,5)を通る二次関数 $f(x)$を求めよ。

１．$f(x)=x^2+x+3$
２．$f(x)=2x^2+3$
３．$f(x)=3x^2-x+3$
４．$f(x)=4x^2-2x+3$

(2)　二次関数 $y=2x^2-5x+5$ の範囲が$-3 \leqq x \leqq 4$ のとき，y の最小値として正しいものを選べ。

１．35／16
２．15／8
３．17
４．38

１１．　半径 7 の円 P と半径 5 の円 Q が異なる 2 点で交わり，二つの円の中心間の距離を d とする。このとき，d の取りうる値の範囲を求めよ。

１．$d < 2$，$12 < d$
２．$2 < d < 12$
３．$d \leqq 2$，$12 \leqq d$
４．$2 \leqq d \leqq 12$

１２．　□に当てはまるものを選べ。

x,y を実数としたとき, $x>1$ かつ $y>2$ は, $xy>2$ かつ $x+y>3$ であるための□。

１．必要条件である

２．十分条件である

３．必要十分条件である

４．必要条件でも十分条件でもない

１３．　$x^2+5x-7=0$ のとき，解はどのようになるか以下の中から選べ。

１．　異なる 2 つの実数解

２．　重解

３．　異なる 2 つの虚数解

４．　1 つの実数解と 1 つの虚数解

【英語】

１４． 下線部の発音がほかの三つと異なるものを，１～４の中から一つ選べ。

１．m<u>ou</u>se

２．l<u>ou</u>nge

３．c<u>ou</u>rage

４．p<u>ou</u>nd

１５． “diploma”の単語の意味を選べ。

１．交渉術

２．卒業証書

３．古門書学

４．外交官

１６． “issue”の意味として**<u>正しくないもの</u>**を選べ。

１．発行する

２．問題

３．(雑誌などの)号

４．学説

１７． 以下の説明が示すものを選べ。

a shop where you can buy food, alcohol, magazines etc, that is often open 24 hours each day

１．Park

２．Convenience store

３．City hall

４．Fire Station

１８． 以下の説明が示すものを選べ。

a large park with many special machines that you can ride on, such as roller coasters and merry-go-rounds

１．Museum
２．Amusement park
３．Movie theater
４．Aquarium

１９． 以下の文章の下線部と同じ意味を表すものを選べ。

Don’t lose your temper. It won’t help you.

１．catch cold
２．be sad
３．get angry
４．feel disappointed

２０． (　)に当てはまるものを選べ。

You should not depend (　　) your parents.

１．in
２．at
３．on
４．to

２１． (　)に当てはまるものを選べ。

(　　) is the host of Tokyo Disney Resort.

１．Leonardo DiCaprio
２．Mickey Mouse
３．Harrison Ford
４．Simon Baker

２２．（　）に当てはまるものを選べ。

Alan Menken (　　) is a composer won an Academy Award for the Original Music Score of “Beauty and the Beast”.

１．whom
２．which
３．what
４．who

２３．(　) に当てはまるものを選べ。

My mother allowed　(　　) to go to the Magic Kingdom with my friends.

１．I
２．my
３．me
４．mine

２４．以下の会話を完成させよ。

A : I persuaded him to buy a new wallet.
B : Wow! How did you do?
A : (　　)
B : That`s awful!

１．I broke his wallet.
２．It was 15,000 yen.
３．He used all of his money.
４．His wallet was terrible.

２５．以下の会話を成立させよ。

A : Please call me Ann.
B : (　　)

１．You should go strait.
２．Don’t you have a cell-phone?
３．OK, Ann. That’s right?
４．OK, I will call.

解　答　用　紙

学籍番号 _____

氏名 _____

【 国 語 】

		3	2	1
8	7	6	5	4

【 数 学 】

9	１０(1)	１０(2)	１１
１２	１３		

【 英 語 】

１４	１５	１６	１７
１８	１９	２０	２１
２２	２３	２４	２５

試験問題等で気が付いたことを裏面に記入してください。

試験問題等について，気が付いたことや不明な点をできるだけたくさん記入してください。試験問題についての記述は，問題番号も記すようにしてください。

謝辞

本論文の執筆にあたって，多くの方のご指導とお力添えをいただきました。ここに深くお礼を申し上げます。

特に，現在の指導教員の石井秀宗先生には，本論文の執筆にあたりご指導をいただきました。研究の大枠から細かい文章表現に至るまで，様々な面において丁寧なご指導を受け賜りました。

また，野口裕之先生には，第 V 実験から始まり先生が退官されるまで，長きにわたりご指導いただきました。先生の下で IRT に関して深く考える機会を賜り，テスト研究をする上での姿勢などを学ばせていただきました。大変感謝しております。

高井次郎先生，光永悠彦先生には，お忙しい中，本論文の審査を担当していただきました。特に論文の構成や研究の発展可能性など，幅広い視野からのご指摘をいただき，本研究の意義を更に深めることができました。

横浜市立大学の山田剛史先生には，テスト作成に関する実践的な機会をいただきました。南山大学の浦上昌則先生には，テストを研究テーマとする最初のきっかけをいただきました。お二人の先生方から学ばせていただいたことは，本研究を進める上で支えとなり今に活きております。

そして，名古屋大学教育発達科学研究科の同期，後輩たち。特に寺尾尚大さんとは，研究のことやそれ以外のことも多く語り合い，同領域の研究者として多くの刺激をいただきました。

最後に，長いモラトリアムを認め支えてくれた家族，一番近くで支えてくれた夫，オキシトシン分泌要員としてモフモフさせてくれた愛犬のこむぎ。みなさまに感謝いたします。

大変ありがとうございました。

令和 4 年 2 月

坪田　彩乃