米国における標準化された試験 －SAT, TOEFL－

独立行政法人 大学入試センター研究開発部　椎名久美子

1. はじめに

本稿では，標準化された公的試験の事例として，米国の大学入試で用いられている SAT と TOEFL を取り上げる．SAT，TOEFL は，いずれも年に複数回実施されているが，得点が標準化されているため，各実施回の得点を互いに比較することが可能である．標準化された得点を用いるテストでは，問題項目（アイテム）の作成や管理，問題冊子の編集，事前テスト（プリテスト）の実施，問題冊子の公開など，テストにかかわるすべての段階において，標準化を可能にするための仕組みを意識した設計が求められる．本稿では，SAT や TOEFL の標準化の仕組みに加えて，標準化を支える各作業段階についても説明する．

米国では，教育カリキュラムが地域や学校によって異なるため，大学入学者選抜の際の学力の指標として，Educational Testing Service（ETS）が実施する SAT（Scholastic Assessment Test）が用いられている（American College Testing Inc.が実施する ACT Assessment も用いられているが，本稿では割愛する）．また，英語を母国語としない者が大学や大学院に入学する際には，TOEFL（Test of English as a Foreign Language）を受験することで英語力が測定される．

SAT には，推論力テスト（SAT I）と科目別テスト（SAT II）がある．SAT I は更に，言語セクション（Verbal Section）と数学セクション（Mathematical Section）に分かれており，前者は言語的推論力を，後者は数学的推論力を測ることを意図して作成されている．SAT II は，英語，歴史と社会，数学，理科，語学などの科目別の学習達成度を測定するための試験である．SAT は年 7 回実施されているが，SAT II の受験機会は，科目によって年 1 回～6 回の幅がある．SAT についての詳細は，藤井他（2001）及び公式サイト(http://www.colleageboard.com/)を参照されたい．

TOEFL は，現在，ペーパーテスト方式（以下，紙筆版と呼ぶ）と Computer-Based Testing 方式（以下，CBT 版と呼ぶ）の 2 種類の方式が実施されている．紙筆版は，1976 年度以降，リスニング（Listening Comprehension），文法（Structure and Written Expression），語彙・文章読解力（Vocabulary and Reading Comprehension）の 3 セクションで構成されている．1998 年から実施されている CBT 版は，リスニング（Listening），文法（Structure），読解（Reading），記述（Writing）の 4 セクションで構成されている．紙筆版は年 6 回，CBT 版はテストセンターに予約して個別ブースで実施される．

2. 標準化の仕組み

2.1 SAT の標準化

SAT は，独自項目と共通項目で構成されている．表 1 に，SAT の旧冊子（実施済冊子）と新冊子が 1 種類ずつ存在し，旧冊子と新冊子が異なる回に実施された場合の例を示す．X は旧冊子のみで出題される項目（独自項目），Y は新冊子のみで出題される項目，Z は新旧冊子で共通の項目（共通項目）である．共通項目は，標準化のためのデータを取得する

ために出題される項目で，採点対象外である．ただし，受験者がどの項目も同じ条件で解答したデータが必要なので，どの項目が採点対象外かは判別がつかないように編集してある．

旧冊子受験者は新冊子の独自項目を受験しないし，新冊子受験者は旧冊子の独自項目を受験しないので，表 1 の Y_old と X_new の部分の成績データは実際には得られない．各冊子は，なるべく同等になるように編集されてはいるが，それでも冊子間の難易度の差がまったくないわけではない．また，旧冊子受験者群と新冊子受験者群は，異なる実施回に受験しているため，同等の学力集団とはみなせない．つまり，データとして得られた X_old と Y_new の差には，冊子間の難易度の差と集団間の学力差という 2 つの差が含まれていることになる．

SAT では，新旧冊子の独自項目の部分の成績（X_old と Y_new）を比較可能にするために，新旧冊子の共通項目のデータを用いて，冊子間の難易度の差（横方向）と集団間の学力差（縦方向）を分離する．共通項目は，内容面や問題の種類という観点から，テスト全体を代表するように出題されており，新・旧冊子受験者群の両方が受験するので，Z_new と Z_old の差が集団間の学力差を反映していることになる．こうして評価された集団間の学力差を手がかりに，仮に新冊子受験者が旧冊子の独自項目の部分を受験した場合の成績 X_new を推定する．

表 1：旧冊子と新冊子の構成例

	独自項目		共通項目
旧冊子受験者群	X_old	Y_old （欠測）	Z_old
新冊子受験者群	X_new （欠測）	Y_new	Z_new

実際の SAT I は 3 時間の試験で，問題冊子は以下のように 30 分または 15 分ずつのセクション単位に分かれている．

- 採点対象
  - 言語(Verbal)セクション：30 分×2 セクション，15 分×1 セクション
  - 数学(Mathematics)セクション：30 分×2 セクション，15 分×1 セクション
- 採点対象外
  - 言語または数学のいずれかのセクション：30 分×1 セクション

配布される冊子に含まれる 7 セクションのうち，採点対象となるのは 6 セクションである．つまり，表 1 の独自項目に相当する部分が 6 セクションを占めている．採点対象外の 1 セクションの中には，表 1 の共通項目に相当する項目と，将来の出題のための事前テスト項目が含まれている．表 2 に，SAT I の問題冊子の構成を示す．採点対象外セクションにおける標準化用の共通項目と事前テスト項目の内訳は，1：2 から 2：3 くらいの割合になっている．実施する際には，どれが採点対象外セクションかが判別がつかないように，まず 30 分ずつの 5 つのセクションを実施して，その後で，15 分ずつの 2 セクションを実施する．採点対象外セクションの挿入位置は問題冊子によって異なるため，受験生には区別がつかない．ただし，採点対象外セクションは言語か数学かのいずれか 1 セクションが割り

当てられるので，30 分単位のセクションについては，言語 2 セクションと数学 3 セクションを受験する者と，言語 3 セクションと数学 2 セクションを受験する者が存在することになる．

表 2：SAT I の問題冊子構成

<table>
<tr><td></td><td colspan="2">採点対象</td><td colspan="3">採点対象外（1 セクション）
言語または数学</td></tr>
<tr><td>旧冊子受験者群</td><td>6 セクション</td><td></td><td rowspan="2">共通
項目</td><td>事前テスト
項目</td><td></td></tr>
<tr><td>新冊子受験者群</td><td></td><td>6 セクション</td><td></td><td></td><td>事前テスト
項目</td></tr>
</table>

SAT II の問題冊子には，標準化のための共通項目が 30～40%くらいの割合で含まれており，この部分のデータを手がかりに標準化が行われている．

SAT I の各セクションと SAT II の科目テストについては，尺度得点の範囲は 200～800 となっている．SAT I の尺度については，1995 年に改定が行われ，1995 年当時の受験者集団の平均を 500 としている．

2.2 SAT の問題冊子の編集

SAT では，前述したような標準化を行うことで，冊子間での得点比較が可能になっているが，内容と難易度がなるべく同じになるよう冊子を作成することも重要である．そのために，SAT I の採点対象外セクションに事前テスト項目を入れて，あらかじめ難易度についてのデータを取得している．事前テストをしてみると，約 2 割は本番のテストには使えない項目であることがわかるという．

事前テスト時の通過率（正答率）は，10%刻みで 10 段階に分類される．SAT I の問題冊子を編集する際に，内容の仕様に加えて，通過率の各段階から何問ずつ集めて問題冊子を編集するかという難易度の仕様が決まっている．その仕様に従って，それまでに蓄積された項目プールの中から自動的に問題が選ばれ，問題冊子が編集される．項目プールには，言語問題は数千題，数学問題は 1,200 題ほどが貯められている．人間が冊子の中身に目を通すのは，冊子が自動生成された後である．このとき十分な数の問題冊子が自動的に生成されないとすれば，それは項目プールが健全でないこと（=内容や難易度の偏り）の証とされる．冊子の編集時に，同一セクション内（言語セクション，数学セクション内）での題材のオーバーラップはチェックするが，言語セクションと数学セクションの間のオーバーラップはチェックされない（人手不足のため）．

SAT I では，常時 14～16 種類の問題冊子が存在するように問題の開発が進められている．3 年がかりで項目を貯めて事前テストを行い，1 年間に新しい冊子を 6 種類作成する．

SAT II は科目別のテストなので，受験生の数が少ない．よって，SAT I のように実際のテストに混ぜて事前テストを実施することが困難である．SAT II の場合は，なるべく実際のテストに近いデータを集めるために，大学の先生などに依頼して，講義の期末テストや組分けテストのような形式で実施してもらっている．高校生，大学生それぞれ 400～500 名によって事前テストを行い，1 問につき最低 400 名分のデータが集められる．生徒には報酬は支払われない．SAT II は，SAT I ほど“High Stakes”な試験ではないので漏洩の

心配は少ないが，それでもセキュリティを高めるため，同じ学校には2度と頼まないようにしている．また，事前テストから実際に出題されるまでの期間をランダムにすることで，万一漏れても，どの実施回に出題されるかの予測を防げるようにしている．

2.3 TOEFL -True Score Equatingによる標準化-

ここでは，紙筆版 TOEFL の標準化の仕組みを説明する．紙筆版 TOEFL では 1978 年以来，項目反応理論（Item Response Theory）を用いた真値による等化（True Score Equating）によって，各冊子の得点を，基準となる冊子の得点へ調整している．

項目反応理論では，テスト（冊子）を構成する各項目に正答する確率を受験者の能力値の関数（項目特性曲線）として表すことで，項目の難易度や識別力が表現される．あらかじめ項目特性曲線がわかっていれば，テストを構成する各項目の特性曲線の和をとって作成されるテスト特性曲線によって，受験者の能力値とテスト得点の関係が表現される．テスト特性曲線がわかっていれば，ある冊子である得点をとることが期待される能力を持った受験者が，もし他の冊子を受けたら何点とれるか，を対応づけることが出来る

ここで，新しい冊子（n 種類）を i=1,2,…,n で表し，それぞれのテスト特性曲線（各冊子に含まれる項目の項目特性曲線の和）を $T_i(\theta)$ とする．なお，これらの項目の特性曲線（項目パラメタ）は，事前テストにより，既知であるとする．また，基準となる冊子（base form）のテスト特性曲線を $T_0(\theta)$とする． True Score Equating は

$$x_{0i} = T_0(T_i^{-1}(x_i))$$

という形で第 i 番目の冊子の得点 x_i を基準となる冊子の得点 x_{0i} へ調整する．ただし，$T_i^{-1}(x_i)$は $T_i(\theta)$ の逆関数である．この形から明らかなように，受験生個人の θ を求める必要はない．図 1 に，基準となる冊子のテスト特性曲線 $T_0(\theta)$と，1 種類の新冊子のテスト特性曲線 $T_1(\theta)$の例を示す．図に示されるように，新冊子の得点 x_0は，その能力値に相当する基準冊子の得点 x_{01}に対応づけられる．

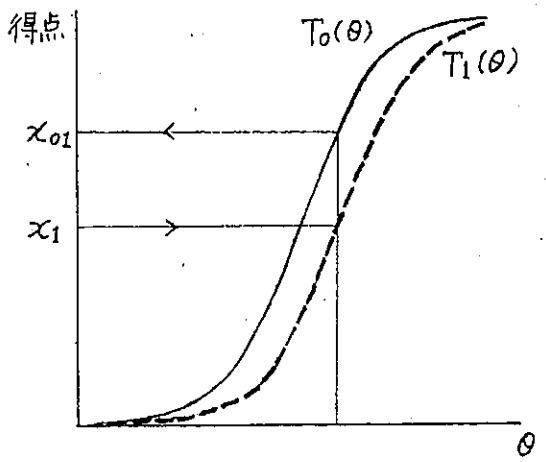

図 1：基準冊子と新冊子のテスト特性曲線

図 2 に紙筆版 TOEFL における標準化のための問題冊子編集の概念図を示す．図 2 は，各実施回の各冊子がパラメタ既知の項目セット（x），事前テスト項目セット（p），新規項目セット（n）で構成されていることを，簡略化して示したものである．また，項目セッ

トは，1から30まであるものとする．

実施回1は，パラメタ既知の項目2セット（x1とx2）のみから構成される冊子が1種類（Y1F1-1）と，それぞれ異なる事前テスト項目セット（p3～p7）が加わった冊子が5種類（Y1F1-2～Y1F1-6）用いられていることを示したものである．事前テスト項目が付加されるのは，北米会場で実施される冊子のみである，Y1F1-1は，海外会場で用いられる冊子である．Y1F1-2～Y1F1-6の冊子が実施されることにより，p3～p7の項目パラメタは既知となり，以後は，x3～x7となって用いられることになる．

実施回2は，パラメタ既知のx3とx4が用いられる冊子（Y2F1系）と，x5とx6が用いられる冊子（Y2F2系）が用いられていることを示したものである．事前テスト項目セットとして，Y2F1系にはp8～p10のいずれか1セット，Y2F2系にはp10～p12のいずれか1セットが加わった冊子が作成される．また，Y2F1系とY2F2系のすべての冊子に，新規項目セットn13とn14が加わる．新規項目セットは，パラメタが未知である点では事前テストセット（p）と同じだが，採点に用いられる点が異なる．Y2F1-1～Y2F1-4，Y2F2-1～Y2F2-4の冊子が実施されることにより，p8～p12，n13～n15の項目パラメタが既知となり，以後はx8～x12，x13～x15となって用いられる．

実施回3は，x8～x11，x13～x14のうちの2セットが，それぞれの冊子で項目パラメタ既知セットとして用いられていることを示したものである．既知セットの組み合わせにより，4種類の冊子系列（Y3F1系，Y3F2系，Y3F1A系，Y3F2A系）が作成される．それぞれの系列について，事前テスト項目としてp15～p20のいずれか1セット，新規項目としてn21とn22が付加されている．実施回3の後では，p15～p20，n21～n22の項目パラメタが既知となり，以後はx15～x20，x21～x22となって用いられる．

実施回4も，これまでと同様，パラメタ既知のx15とx16，x17とx18が用いられる2種類の冊子系列（Y4F1系，Y4F2系）が用いられ，事前テスト項目セットとしてp23～p28のいずれか，新規項目セットとしてn29とn30が付加されていることを示したものである．

実際には，各冊子には半数ほどの事前テスト済みの項目（パラメタ既知項目）と新規項目，それに，米国会場用には事前テスト項目が含まれている．冊子内で，最低40%はパラメタ既知項目が必要である．

各冊子ごとにIRTのパラメタを推定し，その後，事前テスト済みの項目のパラメタを用いてそれらをbase formのスケールにのせる（linking）．これによって，各冊子と基準冊子のテスト特性曲線が比較可能になり，同じ図の中に描かれることになる．

その後，True Score Equatingを行い，各冊子の素点（正答数）をbase formの素点（正答数）に調整する（adjustment）．（図1参照）

最後にbase form用の尺度化の表（変換表）を用いて変換を行う．すなわち，base formの正答数と，尺度点にはあらかじめ対応表が出来ているので，各冊子の素点は，base formの素点を介して，尺度点に換算される．

実際には，紙筆版TOEFLでは，3つのセクションごとにTrue Score Equatingによる標準化が行われる．各セクションごとに正解数が尺度得点に換算され（得点範囲は31～68または31～67），テスト全体の得点は，各セクションの尺度得点を加算した値に(10／3)をかけることで算出される（得点範囲は310～677）．

x　項目パラメタ既知（p やｎだった項目が実施を経てパラメタ既知となったもの）
p　事前テスト項目
n　新規項目（採点対象）

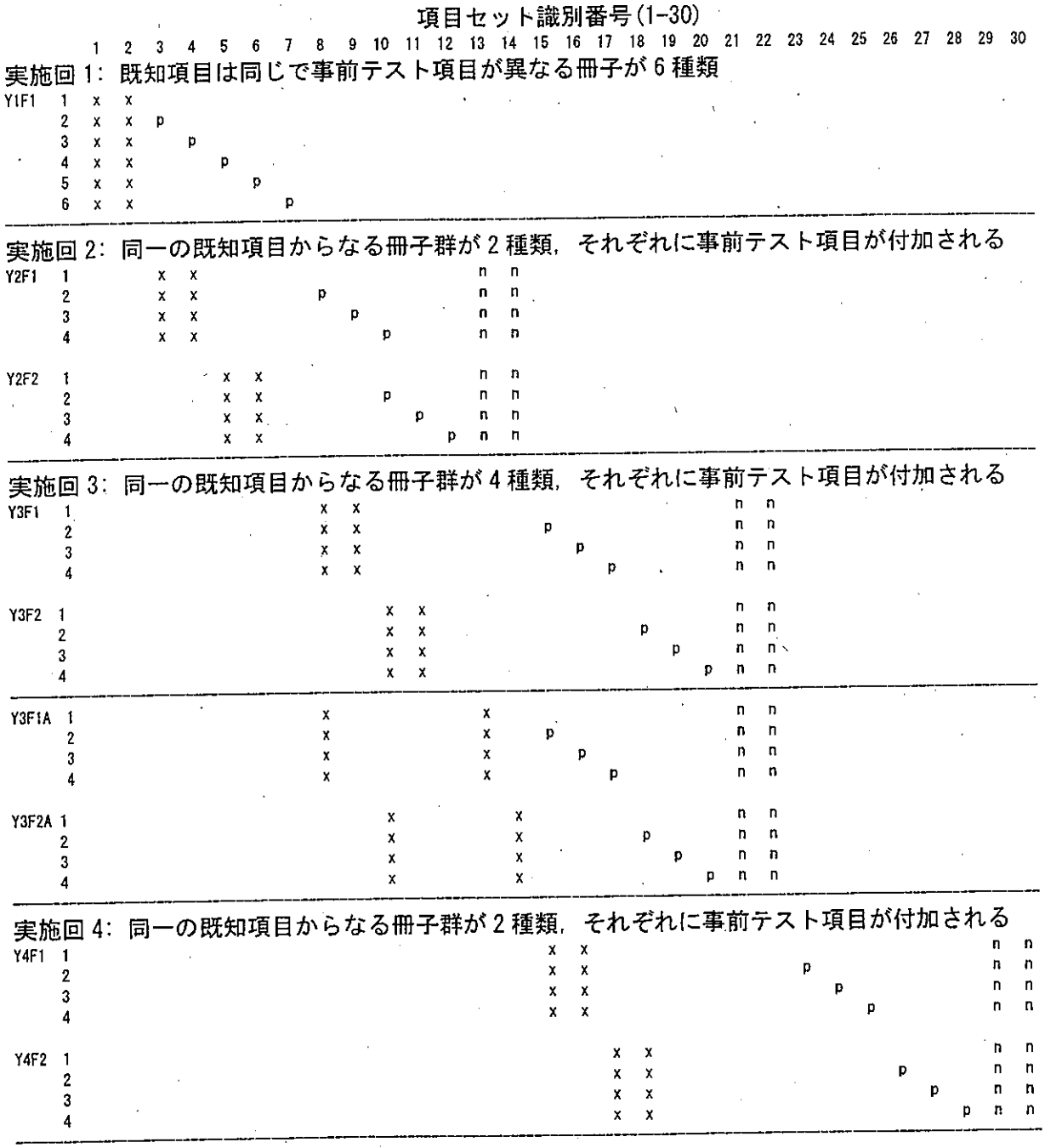

図 2：紙筆版 TOEFL における問題冊子編集の概念図

2.4 紙筆版 TOEFL の冊子編集と事前テスト

事前テストの実施は，項目パラメタを事前に知るために不可欠であるが，セキュリティを確保しつつ質のよいデータを収集するために，冊子の編集に様々な工夫がなされている．

紙筆版 TOEFL では，年に 12 種類の冊子が実施されるが，そのうち北米（合衆国とカナダ）会場で実施される 6 種類には，各セクションに事前テスト項目が混ぜ込まれている．そのため，国外よりもテスト実施時間が全体で 40 分長くなる．事前テストが北米のみで実施される理由は，実施回ごとの受験者の質が安定しているためである（各回 12,000〜40,000 人が受験）．

事前テスト項目は，採点対象の問題の間に埋め込まれており，受験者には判別がつかないようになっている．事前テストは，事前テスト項目を混ぜた冊子を 15〜20 種類用意して，約 5,000 人にランダムに配布することで実施され，事前テスト項目 1 問につき約 500 人のデータを集める．

一度事前テストされた項目は，平均して 6〜8 ヶ月の間をおいて，1 回だけ採点対象項目（operational items）として出題される．項目プールには数千題の項目が常時蓄積されているが，プールの中に置いておくのは 7 年間と決まっており，その期間を越えるとプールから削除される．TOEFL は 1 年に 6 回実施されるので，事前テスト後に採点対象として同じ項目が再び出題されるのは，6 回／年×7 年 ＝42 回のうちのどれかの回ということで，予測は困難であると考えられている．

TOEFL においても，受験者が問題冊子を持ち帰ることは禁止されているが，制度的に非公開にするだけでなく，それを補うための仕組みが確立されている．多種類の問題冊子の中にプリテスト項目を埋め込んだり，採点対象項目として出題されるまでの期間を予測されないようにしたりすることは，事前テスト項目が漏れたとしても被害を最小限に食い止めるための仕組みである．と同時に，事前テスト項目を盗み出す労力に見合った結果が得られないことを暗示することで，受験者が項目を盗もうとする意欲を抑制する仕組みとも言える．それでも，項目を不正に入手しようとする組織的な活動がいくつか確認されている．ETS には法律の専門家もおり，項目を盗んだ相手に対しては裁判に訴える場合もあるという．

2.5 CBT 版 TOEFL について

CBT 版 TOEFL の外見上の特徴は，個別ブース内の端末画面上に問題が提示され，マウスを使って解答するという点であるが，一部のセクションが適応型テスト（Adaptive Test）になっている点も大きな特徴の一つである．

紙筆版 TOEFL のテスト冊子には，易しい項目から難しい項目まで，まんべんなく含まれており，難易度という点でも，また，内容的にも，冊子間の違いがなるべく小さくなることを目指して編集されている．受験者は，たくさんの似たような冊子の中のどれかを受けることになる．

それに対して，適応型テストでは，最初に中くらいの難易度の項目が出題された後は，受験者の正誤反応に応じて，項目プールの中から，その受験者の能力を推定するのに最適と判断された項目が出題されていく．受験者の能力は，項目特性曲線が既知の項目の情報を用いて，随時推定される．受験者が正解すれば，次に出題される項目はだんだん難しく

なるし，間違えば，だんだん易しくなる．また，正解数が同じでも，難しい問題に多く正解した受験者のほうが，得点が高くなる．

適応型テストの場合は，受験者の正誤に応じて次に出題される項目が選ばれるので，一時に提示されるのは1問であり，しかも，スキップすることは許されない．よって，CBT版 TOEFL では，4つのセクションのうち，リスニングと文法が適応型テストとなっている．読解セクションは，1つの文章につき複数の問題項目があらかじめ決められているので，端末上で実施されてはいるが，テスト方式としては従来の紙筆式と同様，様々な難易度の問題が含まれている．また，記述セクションも，与えられた題材についてのエッセイを書く方式で，適応型ではない．

記述セクション以外はコンピュータが採点を行うので，受験者はテスト終了直後に，自分のだいたいの得点を知ることが出来る．CBT版 TOEFL の得点範囲は 0～300 で，紙筆版 TOEFL の得点範囲（310～677）と，まったく重ならないように設定することで，CBT版と紙筆版の得点の混同を防いでいる．

CBT版 TOEFL でも，紙筆版と同様に，各セクションの出題項目数はあらかじめ決まっており，項目特性曲線が既知の項目と事前テストのための項目（採点対象外項目）が，受験者に区別がつかないように混ぜ込まれて出題されている．事前テスト項目は，約 25～35%を占める．

CBT版では，受験者ごとに異なる問題項目のセットを出題されるので，紙筆版のように，採点対象項目として1回出題しただけで廃棄していたのでは，すぐに問題項目が底をついてしまう．よって，出題回数がある数を超えたら，以後出題しないようにすることで，必要な問題項目数を保っている．それでも，紙筆版よりは，多くの問題項目が必要であることに変わりはないので，問題項目の開発の負担は従来より大きい．また，テストの実施に必要なコンピュータを備えたテストセンターを開設・維持する費用も大きく，ETS では，採算の合わないテストセンターは，紙筆版 TOEFL に戻す方向で，見直しを行っている．

3. 問題項目作成の流れ

試験を実施するためには，試験問題（問題項目）が必要となるが，問題項目作成の流れに着目してみると，標準化された試験と日本の従来型の試験（非標準化試験）とで大きく異なる点がある．非標準型の試験では，多くの場合，特定の試験実施回を念頭において，試験に使用する問題冊子に必要な問題項目が作成される．それに対して，米国における標準化試験の通常のプロセスでは，多数の項目ライターから集められた問題原案が，改良段階を経て項目プールに蓄積される．任意の実施回の問題冊子の編集にあたっては，蓄積しておいた多数の問題項目群の中から難易度や内容などのデータに基づいて適当と思われる項目が選択されることになる．また，前述したように，実際に（採点対象として）出題されるのに先立って，問題項目の特性に関するデータが何らかの形で収集されて（事前テスト），問題冊子の編集や標準化手続きのために用いられる点も，標準化テストの大きな特徴である．

問題項目の作成は，外部の項目ライターと ETS 内部のレビューアによって行われている．例えば，SAT I や TOEFL の読解（reading）に関する問題（SAT I や TOEFL）の場合，項目ライターがはじめから完成品を提出するのではなく，まず題材となる文章（passages）

を項目ライターから幅広く集めてから，ETS 内部の項目開発者（レビューアなど）が外部の項目ライターをコントロールして，1 問に仕上げていく流れになっている．

3.1 SAT の場合

SAT の場合は，問題の種類や科目によって項目の作成方法が異なっており，以下の 3 種類が挙げられる．

1) 外部ライターと内部開発者で項目を作成するテスト（SAT I 数学セクション，SAT I 言語セクションの中の「読解」タイプ）
2) 内部開発者のみで項目を作成するテスト（SAT I 言語セクションの「アナロジー」タイプ及び SAT I 数学セクションの「量の比較」タイプ）
3) 委員会形式で項目を作成するテスト．年に 2 回，ETS に委員が集まり，作成してきた問題について 2 日半のレビューを行う．（SAT II の一部の科目）

SAT I の「読解」に使われる文章は，アメリカの様々な地域の項目ライター（10～12 人）から集められるが，文章が貧弱な場合であっても却下せずに，ETS 内部の開発者が編集して手を加えることによって，項目として仕上げていく．

SAT I における ETS 内部の開発者の数は，SAT I 言語問題が 15 人，数学問題が 12 人ほどである．SAT II の生物やフランス語などの科目については，内部開発者は全科目合わせて 400 人ほどである．

3.2 TOEFL の場合

図 3 は，紙筆版 TOEFL の読解及び Structure 問題における，項目作成の流れ図である．

外部の項目ライター（Outside Item Writer）
↓
項目の作者（Item Author）
↓
第 1 レビューア（1st reviewer）
↓
第 2 レビューア（2nd reviewer）→ 公平性のレビューア（Fairness reviewer）
↓
決定者（Resolver）← 編集者（Editor）
↓
ロック・レビュー（Lock Review）
（最終レビュー：答がただ 1 つかどうかなど，最終的にチェックして凍結）
↓
事前テスト段階へ（以後は問題に手を加えない）

図 3：紙筆版 TOEFL の"Reading"及び"Structure"の項目作成の流れ図

図 3 の流れのうち，大半の時間は，項目ライターが問題に使用する文章を見つける時間にさかれる．TOEFL の読解問題の場合，項目ライターは問題の文章として適当と思われるもの（大学の最初の 2 年で読むようなもの）を探して，まず文章だけを ETS に送る．この時点では設問項目はまだ作成しない．ETS の内部には，文章を読んで採否を判断するテスト項目開発者が 25 人ほどいる．内部の開発者は文章を読むだけで，設問は作らない．

図3中の第1及び第2レビューアが内部の開発者であり，文章の採否を決定する作業の他に，外部の項目ライターにコメントを返して文章を改良させる作業も行う．読解，聴解，作文技術の3つのセクションごとにリーダーがいて，項目ライターと連携をとる役割を果たしているが，聴解セクションの内部開発者は読解よりももっと少ない人数である．項目を1つ作成するには，平均して18ヶ月かかる．

開発者がその文章の善し悪しを判断する段階では，外部の項目ライターから送られた文章のうち，採用されるのは半分ほどである．過度に専門的なものや，読者が強い思い入れを持つような文章は，この段階で除かれる．文章が採用になって，項目ライターのもとに送り返された後で，1つの文章につき5〜6問の設問項目が作成され始めることになる．TOEFLのように必要とされる項目の数が多いテストでは，似たような話題の文章が異なる項目ライターから提出された場合でも，テストを実施する時に同じ冊子に入りさえしなければよいので，特に話題の重複を理由として作成をやめさせたりしない．

公平性のレビューア（Fairness Reviewer）と編集者（Editor）の役割は，TOEFL以外の部署の人が行う．公平性に関するレビューは，民族や性による偏見やステレオタイプ的な見方が含まれていないかどうかをガイドラインに沿って中身をチェックするものである．また，編集者は，英語として間違いがないかをチェックする．編集者が語を差し替えると問題の難しさも変わるので，決定者が語の変更点などの最終調整を行う．その後は，答がただ1つかどうかなど，最終的にチェックするためのロックレビュー（Lock Review）を行い，以後は問題項目に手を加えずに事前テスト段階に進むことになる．

なお，ETS内部における項目開発者の役割は固定されたものではなく，問題セットごとにレビューアと決定者の役割を入れ替えている．

CBT版TOEFLの場合は，どこをクリックさせるかなどの情報を問題画面に埋め込む役割の人（Formatter）が図3の流れの中に加わることになる．

4. 項目ライター制度

「項目ライター」という言葉は日本では耳慣れないが，アメリカでは1つのキャリアとして認知されている．また，標準化試験の遂行を可能にするに足るだけの多数の項目を常に作り続けるために，能力の高い項目ライターを育成するための訓練も重視されている．

ただし，組織の外部のライターが作題することで，問題が漏洩する危険が存在することも事実である．ETSは，1人の項目ライターが作成した項目が1つの問題冊子に占める割合を非常に小さくしたり（SAT I），項目の作成から出題までの期間をランダムにして予測を難しくしたりすることで（TOEFL），漏洩のリスクを小さくできるとの見解を示している．

4.1 SATの項目ライターとその育成

3節で述べたように，テスト開発は外部の項目ライターと内部のテスト開発者・項目ライターの連携によって進められていることが多いが（テストによっては委員会形式），内部と外部の人数比は，テストの種類によって異なる．SAT IIは，6人の委員からなる委員会で問題項目が作成されるが，作成や手直しはすべて委員会で行われる．6人中3人が高校の教師で，残り3人はそれ以外の人と決められている．

外部の項目ライターは高校の教師や元ETS勤務者などが多い．SAT I言語問題の場合は，元 ETS 勤務者が主な外部の項目ライターであるのに対して，SAT I 数学問題の場合は，高校の教師が外部の項目ライターの主力である．項目ライターには，作成した問題項目の採否にかかわらず，1 問あたりいくらで作成報酬が支払われるが，額としては多くはない．また，SAT I 数学問題では，1 人が作成する項目数が多過ぎると似たような内容ばかりになるので，1 人 15 問くらいを適切な作成数としている．

ETS では，SAT I の項目ライター育成のため，1 年に 1 日，ワークショップを開催している．ワークショップは，10～15 人程の高校の教師をカレッジに集めて行われる．参加者にはガイドラインと例題を事前に渡して，8 問作成してワークショップに持って来るように言っておく．ワークショップでは，午前にその 8 問を見て講評して，午後はグループに分かれて，数題については講評しあって，完成させる．参加者は，自宅に帰ってから 8 問を完成させた後，更に 15 問作成する．参加費用（1 日あたり 300 ドルと交通費）は ETS が負担する．ETS では，1，2 年おきにワークショップを開くことで，同じライターを繋ぎとめておけるだろうと考えている．

外部の項目ライターの任期には，特に決まりはない．長い間続ける人もいれば，すぐやめてしまう人もおり，様々である．項目作成の報酬は，20 年前は 10 問で 100 ドル，現在でも 10 問で 300 ドルほどであり，高くはない．項目ライターは，報酬目的というより，一種のステイタスととらえられているようだ．

4.2 TOEFL の項目ライターとその育成

TOEFL の場合は，項目ライターの志望者のほうから ETS にコンタクトして来る．他の項目ライターの紹介や，博士号を取得した学生で将来 ETS で働くことを目指す者が多く，ほとんどはニュージャージー州近辺在住者である．（注：ETS はニュージャージー州プリンストンにある．）

志望動機としては，項目ライターが職業的に良いキャリアと考えられているから，というのが多い．ETS からの報酬はそんなに高額ではないが，中にはプロのテスト開発者となる人も少数いる．また，以前 ETS に勤務していた人が良い項目を作成する場合が多い．

育成方法としては，志望者が何人か集まったら ETS に週末来てもらって，基となる文章としてはどのようなものが適当か，どのように問題項目を作成するか，等についての講義をする．講義の最後にテストをして，合格した人には自宅に仕事を送り，項目ライターになってもらう．

5. 問題項目の管理

SAT でも TOEFL でも，標準化を可能にするためには，膨大な数の問題項目の流れを把握し，管理しなければならない．ETS で扱われるすべての問題項目は，共通のデータベースに入れられて管理されている．データベースへの項目の入力は，図 3 の流れで言うと「項目の作者」（Item author）が行う．その後の作業は，すべて項目データベースにアクセスすることで行われる．データベースには，項目の仕様（紙筆テスト用／CBT 版用）や，扱っている話題のカテゴリなどの情報が入っている．どの項目がどの処理段階にあるか，誰がどの段階をどれくらい処理しているかを追いかけたり，扱っている話題をカテゴリ別に

検索したりすることが出来る．

事前テスト後は，統計パラメタに関する情報もデータベースに加えられる．問題冊子作成時には，問題の自動選別システムプログラムが，データベース上の統計パラメタを参照して，仕様に応じた問題冊子を仮編集し，その後で人間がチェックする．

6. 問題の公開

標準化テストにとっては，問題の公開もコントロールしなければならない事項の１つである．事前テストは本番のテストに混ぜ込まれて実施されることが多く，実施された問題冊子をすべて公開してしまうと，次回出題された際に問題の難易度などの性質が変化する可能性が大きく，収集したデータを以後の標準化に用いることが困難になる．幸い米国では，問題の公開に対する要求よりも，問題内容が特定のグループに対する差別や有利不利を含まないことに対する要求が強い傾向があったため，問題冊子を回収することは，受験者にとって不自然なことではなかった風潮がある．

SAT や紙筆版 TOEFL では，本番テストとして使用済みの問題冊子を何種類か集めたものが，公式サンプルテスト集として販売されている．サンプルテスト集では，事前テスト項目は削除されている．SAT の場合は，７つのセクションの中の１つが，標準化のための共通項目と事前テスト項目を含んでいるので，そのセクションを除いたものが公開されることになる．１つのセクションをまるごと除去して公開すればよいので，公開用冊子を改めて編集する労力を省くことが出来る．紙筆版 TOEFL では，それぞれのセクションに事前テスト項目が埋め込まれているので，それらを除いたものを公開用冊子として編集する．

近年は米国においても，問題の公開に対する要求が高まりつつある．ニューヨーク州では，ある程度以上の人数が受けるテストについては，問題を公開せよという法律が制定された．SAT II は，受験人数が少ないので公開しなくてよいが，SAT I は受験人数が多いのでこの法律の適用を受けることになる．ただし，公開対象となるのは，採点対象となる項目のみで，事前テスト項目は公開を免れている．SAT I では，１年間に新しい冊子が６種類作成され，６種類が使用済み扱いとなる（常時 14～16 種類の冊子が存在するように問題開発が進められている）．使用済みとなった６種類のうち，４種類の冊子が公開される．

7. 考察

SAT，TOEFL と標準化された試験について概観してきたが，いずれの試験も，セキュリティを保ちつつ多数の問題項目を作成し続ける制度や，標準化の仕組みに応じた問題冊子の編集及び事前テストの組み合わせによって，標準化が可能になっている．SAT I では事前テストや共通項目のためのセクションが設けられている．TOEFL では特別のセクションを設けないが，問題冊子の中に事前テスト項目や共通項目が埋め込まれる．SAT II では受験者が少ないために事前テストには苦労しているが，それでも可能な限り受験者に近い集団を用いて事前テストを行う努力が払われている．

ここで，試みとして，日本で事前テストを実施する際に考慮するべき点をいくつか挙げてみる．

まず，事前テスト項目が本番の問題冊子の中に存在するということは，問題冊子の持ち帰りを禁止して，試験問題の公開を何らかの形で制限する必要があることを意味する．受

験生が問題冊子を持ち帰るのが当然で，一度出題された問題はすべて公開されるのが原則という風土においては，受験生や受験産業からの抵抗が大きいと考えられる．

次に，事前テスト項目が存在することが公になれば，それらは次回以降に再度出題される可能性の大きい項目として，とらえられるだろう．問題冊子を非公開にしたとしても，すべての問題項目を再現して以降の試験に備えようとする受験産業が出ると予想される．そうなれば，事前テストの際の難易度と，次回出題された際の難易度は大きく変化してしまい，標準化が困難になってしまう．

対策としては，米国で行われているように，多数の項目を作成したり，事前テストと次に出題されるまでの期間をランダムにしたりすることで，多数の項目を収集して正答を覚える労力の結果得られる得点が非常に小さいものになるような試験システムを作成することが考えられる．その場合，同じ実施回に複数種類の問題冊子を用いることで一度になるべく多数の項目の難易度データを収集したり，事前テスト項目が人目に触れる範囲を小さくしたりする工夫が必要であろう．

複数の種類の冊子を用いるという点に関しては，どの種類の冊子に当たっても難易度が調整されて不公平がないことが，受験者やテストの利用者（大学など）に正しく理解される必要がある．そのためには，標準化の仕組みを説明する必要があるが，事前テスト項目の比率や問題冊子の編集及び配布方法について明らかにし過ぎると，事前テストによるデータ収集の仕組みに影響が出てしまう恐れもある．標準化を実現するための仕組み，項目の作成及び管理の体制，問題冊子の編集・配布・公開など，テスト全体を広く見通した設計が求められるだろう．

参考文献

芝祐順編（1991），項目反応理論，東京大学出版会．

藤井光昭・柳井晴夫・荒井克弘編著（2001），大学入試における総合試験の国際比較－我が国の入試改善にむけて－，多賀出版．

Kolen M. (2001), "Test Equating: Purpose and Design", e-Learning Forum 2001 Winter 配布資料．

補足

現在，ETS では，新しい SAT を 2005 年 3 月に実施する予定で改定を進めている．主な変更点は，以下の 3 点である．(http://www.collegeboard.com/about/newsat/newsat.html 参照)

1) 言語セクション(Verbal section)の名称を Critical reading section に変更する．アナロジータイプを削除して，代わりに短文読解タイプが加わる．25 分×2 セクションと 20 分×1 セクション．
2) 新しいセクションとして，Writing section が加わる．このセクションは多肢選

択式の文法問題と小論文(written essay)で構成される．50 分．

3) 数学セクション(Math section)の出題範囲の拡大．現在は高校で 2 年間勉強した範囲だが，3 年間勉強した範囲まで広げる．25 分×2 セクションと 20 分×1 セクション．

それぞれのセクションの得点範囲は 200～800 になることが予告されている．また，25 分の採点対象外セクションの存在が予告されている．

また，TOEFL に関しても，2005 年 9 月に新しいタイプの TOEFL への変更が予定されている．(http://www.ets.org/news/03022501.html)