高大接続改革の技術的基盤　―テスト理論活用の観点から―

南風原朝和（東京大学）

本稿は，現在進行中の高大接続システム改革会議の第 5 回（2015 年 8 月 5 日）時点での「中間まとめ」（案）（文部科学省 Web ページより入手可能）に基づいて，そこで検討されている 2 つの新テストの開発と利用に関する諸問題について，主にテスト理論の観点から検討するものである。

1　実施の目的は何か ― 妥当性のための問い

「テストはその実施の目的に適っているか」という意味での妥当性の問いに答えるには，「そのテストは何のために実施するのか」が明確に示されなければならない。「高等学校基礎学力テスト（仮称）」の目的について，まとめ案には“生徒の学習意欲の喚起，学習の改善を図るとともに，その結果を指導改善等にも生かすことにより，高等学校教育の質の確保・向上を図ることを主たる目的とする”（p.14）とある。一方，科目は主に高校 1 年次で履修する「国語総合」「数学Ⅰ」「コミュニケーション英語Ⅰ」に限られている（p.15）。このテストを 2 年次，3 年次で受検することが，なぜ学習意欲の喚起や学習の改善，指導の改善につながるのか，そのロジックの説明が必要であろう。また，このような目的で実施するテストについて，受検料を自己負担させる（p.20）のであれば，それを正当化する説明が必要であろう。

2　何を測定するのか ― 妥当性のための問い

「テストはそれが測るべきものを正確に測っているか」という意味での妥当性の問いに答えるには，「そのテストで何を測ろうとしているのか」が明確に示されなければならない。「大学入学希望者学力評価テスト（仮称）」について，まとめ案には“十分な知識・技能が習得されていることを前提に，「思考力・判断力・表現力」を中心に評価する”（p.38）との記述がある。この記述からは，前提となる十分な知識・技能が習得されていない場合は何が評価されているのか，という疑問が残る。一方，“「知識・技能」のみならず「思考力・判断力・表現力」を評価する”（p.34）との記述もあり，「知識・技能」の位置づけが明確でない。

「思考力・判断力・表現力」については，「知識・技能」「主体性・多様性・協働性」とともに「確かな学力」の 3 要素の 1 つとされ，「高等学校基礎学力テスト（仮称）」においても評価の対象とされている。しかし，思考力と判断力と表現力はそれぞれ別の能力であり，全体として 1 つの要素とするのは無理があるのではないか。いずれにせよ，これらを科目ごとに評価するのであれば，それぞれ具体的にどのような能力を表わすのかが，サンプル問題とともに明示される必要がある。

なお，B. S. Bloom の教育目標の分類学においては，「知識」「理解」「応用」「分析」「総合」「評価」の 6 カテゴリが設定されている。このうち，「知識」の上位に位置づけられる「理解」については，学校教育法においても，「生活に必要な数量的な関係を正しく理解」などの形で目標として掲げられている。実際の教育においても，学習内容に関する深い理解，本質的な理解，統合的な理解は重要な目標であると考えられるが，2 つの新テストともに，理解の評価についてまったく触れていないのは疑問である。

3　広範な学力の受検者に対応可能か ― 測定精度・情報量に関する問い

「大学入学希望者学力評価テスト（仮称）」については，“広範囲にわたる受検者が受検する可能性があるため，問題の難易度をできるだけ広範囲に設定する”（p.41）とある。これは測定精度・情報量を広範囲にわたって保持することを意味するが，同時に，それぞれの水準での測定精度・情報量は高くならないことになる。“高難度の問題を選択できるようにする”（p.41）との記述もあるが，それぞれの水準での測定精度・情報量を高めるためには，難易度の異なる複数のバージョンを用意することも考えられるのではないか。あるいは，個別入試において学力の評価を相当程度，補強する必要があるのではないか。

4　複数回受検は可能か ― 得点等化に関する問い

複数回受検は，それぞれのテストの目的および結果利用の仕方によって，その必要度が判断される。複数回受検のためには，問題作成および実施のコストが増大するほか，異なる実施回の間での得点の比較可能性が問題となる。上記 3 の難易度の異なる複数バージョンの場合は，共通項目を含めることによってその情報を等化に利用することができるが，複数回受検の場合は共通項目を含めることができないので,別の情報により等化を試みる必要がある。

5　目標準拠・学習診断は可能か ― 結果の解釈に関する問い

「高等学校基礎学力テスト（仮称）」は，“基礎的な学習の達成状況について確認する「目標に準拠した評価」（いわゆる絶対評価）を行う”（p.19）とあるが，それぞれの科目が広い範囲の内容を扱っているなか，どのようにして基準設定を行うのか。また，項目反応理論（Item Response Theory，IRT）を適用すれば“学習内容（又は試験問題）に対する理解度についての絶対評価としても利用可能”（p.19 脚注）とあるが，基本的にテスト問題全体を使って受検者を 1 次元に序列化する方法である IRT にそのような利用が可能なのか疑問である。

6　結果の表示はどうするか ― 尺度に関する問い

以前の議論では「1 点刻みを廃して」という方向性が示されていたが，まとめ案ではそのような表現はなく，“結果の多段階による表示による提供を行うこと，あわせて，種々の具体的なデータ（例えば，パーセンタイル値に基づき算出されたデータ，標準化得点，出題分野ごとの正答数や誤答数など）を大学に提供することなどについて・・・今後より専門的に検討する”（pp.41-42）とされている。段階表示についてはそのロジックと具体的方法，パーセンタイル値や標準化得点のようにそのときの受検者集団の分布に準拠した尺度を用いることについては，そこから派生しうる問題について，検討が必要である。

7　記述式問題，IRT，CBT の導入は可能か ― 実行可能性に関する問い

記述式問題，IRT，CBT（Computer Based Testing）の導入については，その実行可能性（feasibility）に関して，以前にくらべると，より慎重な検討がなされるようになってきている。記述式問題と CBT については，数十万人という受検者規模が重要なファクターであり，IRT については，項目バンクの規模（蓄えておく項目数），項目の公開・非公開の問題，1 次元性の仮定の問題などが考慮・検討されなければならない。特に項目バンクの規模を大きくすることについては，学習内容の「理解」などではなく，「思考力・判断力・表現力」を中心に評価するという高いハードルが問題作成に課せられることからすると，かなりの困難が予想される。そして，IRT の導入が難しいとなると，IRT に基づく等化（とそれを前提とした複数回受検）や，CBT による適応型テストも難しくなる。いずれも,「先に導入ありき」ではなく，実行可能性について一層慎重に，かつ実証的に検討する必要がある。