測定・評価部門

教育評価・心理測定で用いる測度の妥当性検証の機運と
社会的役割を担う試験をめぐる課題解決への取組み

内 田 照 久

(独立行政法人 大学入試センター)

Trends in Examining the Validity of Psychological Scales in Educational Evaluation and Psychological Measurement and Efforts to Solve Problems in Testing With Social Responsibilities

Teruhisa UCHIDA

(NATIONAL CENTER FOR UNIVERSITY ENTRANCE EXAMINATIONS)

Following the trend from previous years, many studies were based on the construction of psychological scales. In contrast, because the field shares the concerns on oversupply and overabundance of psychological scales and measurements, many seminars and research presentations have warned the necessity of validity examination and methodology for it. In recent years, studies of psychological measurement, educational evaluation, and testing are subdivided into several areas and reported in various academic organizations. While some studies are conducted on theories of testing, the development of new statistical procedures, and application of IT technology on testing, others investigate practical problems occurring in real-life testing situations. In the future, these studies may provide new directions to connect academic research and educational practices, which has been a problem to be solved in educational psychology.

Key Words : Psychological Scale, Construct, Validity, Educational Evaluation, Test

教育心理学領域の測定・評価に係わる研究について、2010～2011年の動向を概括した。従前から続く傾向として、心理尺度の作成に係わる研究が数多く進められた。しかし本年は、心理学的な構成概念や測定尺度の過度の氾濫や過剰供給に対する危機感から、尺度の妥当性の検証の必要性を訴えると共に、具体的な検証方法を提案するセミナーや発表が多くなされたのが特徴的であった。また近年、心理測定や教育評価、テスト理論の研究領域は、その発展に伴って細分化が進んでいる。試験やテストに関連した教育評価の研究は、教育心理学会からテスト学会などに発表の場が移動しつつあるように見受けられた。そこではテスト理論や統計手法の開発、IT技術の活用などのテスト技術に関する研究が精力的に進められる一方、社会で実際に運用され、受験者の処遇を左右するテストや試験の場面で発生している具体的な問題に向き合い、その解決を目的とした研究も行なわれている。これらの研究は、教育心理学の従来からの課題でもある研究と教育実践の社会的な繋がりを考える上で、大切な方向性の1つを示すものと考えられる。

キーワード：心理尺度、構成概念、妥当性、教育評価、テスト

1. 心理学的な構成概念の提案・測定尺度の開発

教育心理学領域の測定・評価の動向としては、従前から続いている新たな構成概念の提案、心理尺度の作成改良に係わる研究が、本年も数多く進められてきた。

1.1 心理尺度の作成に関する研究

まず、心理尺度の作成や検討に係わる研究で、『教育心理学研究』に掲載されたものについて概観してみることとする。

小浜 (2010) は、大学生が学業課題などを先延ばしにした時に感じる意識のあり方に着目し、先延ばし意識特性尺度を作成している。そこでは、先延ばしをする前の意識の下位には、否定的感情が伴う決断遅延、楽観視による習慣的な行動の遅延、気分の切り替えを目的とする先延ばし傾向の3種類の性質があるとした。

高坂 (2011) は、Adlerの個人心理学における概念の1つである共同体感覚 (Social Interest) を取り上げて、所属感・信頼感、自己受容、貢献感の3つの下位尺度から成る共同体感覚尺度を作成している。

また教育心理学会の測定・評価のセッションでも、いくつか心理尺度の提案がなされている。

例えば、近藤・金子 (2011) は、母親の育児対処能力に関する尺度を作成している。崔・庄司 (2011) は、中学生を対象とした自己防御 (Self-Regulation) 尺度を提案。また、岩木・梅津 (2011) は、知能感 (Theory of Intelligence) 尺度の日本語版を発表。さらに今野・土肥・宮川 (2011) は、数学学習におけるモチベーション評価尺度を提案している。その他にも、他の多くの領域において、数多くの心理尺度の提案がなされていた。

『心理学研究』にも目を向けてみると、測定尺度に関連したものとしては、下記の研究が上げられよう。

赤田 (2010) は、測定対象者を職種としての保育士に絞り、保育士ストレス評価尺度 (Nursery Teacher's Stress Scale) を作成している。そして、それを用いて、管理職者とそうではない者の、職階によるストレス構造の違いを検討した。

井潤 (2010) は、発達障害のケアを念頭においた上で子どもの養育行動を捉えるため、Parenting Scaleの翻訳を行なった。その日本語版について、探索的因子分析を行なって尺度の検討を試みた。

並川 (2011) は、自分自身の能力やパーソナリティのあり方を考える際に、他者との比較ではなく、過去の自己と現在の自己を対比して評価するという、経時的比較の観点に注目した。そして、その自己の経時的比較をよく行う者と、あまり行なわない者の間の傾向性の違いを測るための、経時的比較志向性尺度を提案した。それにより、自己理論研究の実証的な取組みへの貢献を目指している。

小山内・岡田 (2011) は、読者が文学作品に触れた時に生ずる心理的な体験や情感を測定の対象としている。そのために、物語理解に伴う主観的体験、感情の喚起の状態を多面的に捉えるための文学反応質問紙 (Literacy Response Questionnaire, LRQ) の日本語版 (LRQ-J) を作成した。その分析の結果、原版とはやや異なる、

物語世界への没入、読書への没頭、作者への関心、現実の理解、ストーリー志向といった下位尺度を抽出している。

このように、新たな心理学的な構成概念の提案と、その測定のための心理尺度の作成の研究の数は、枚挙にいとまがなく、全てを網羅的に把握すること自体、困難を伴う状況である。

1.2 既存の心理尺度の再検討・短縮版に係わる研究

これまでに発表されてきた尺度を見直すための検討や、従来版の尺度の短縮版の作成や評価に係わる研究も進められている。

安達 (2010) は、以前、安達自身が作成した職業意識に関連するキャリア探索尺度の再検討を行なっている。旧版は女子学生のための結果から考案されたものであったが、この研究では男女学生を対象として再検討した。その結果、尺度そのものは妥当な測度であるとしながらも、下位構造での性差の存在については検討する余地があるとしている。

谷・並川・脇田・中根・野口 (2011) は、子どもを対象とした特性不安尺度であるところの日本語版 State-Trait Anxiety Inventory for Children (STAI-C) の短縮版の作成を試みている。そこでは、項目応答理論 (IRT) の分析結果に基づく項目選択を行なっている。

鋤柄 (2011) は、性格検査のMMPI、および Big Five の尺度の短縮版の評価を行なっている。短縮版の項目は、完全版の項目の一部であるため、完全版と短縮版での単純な相関をもとめると、同一の項目が重複しているためにその値は過大評価される。その問題点の克服のためにマッチング法が考案されているが、その有効性を検証するために、シミュレーションデータを用いた検討を行なった。

このように、既存尺度の改良に向けた継続的な検証、実用面での要請をふまえた実施の簡便化を図るための短縮版の作成や評価に関する研究も、相対的にはまだ数は少ないものの着実に進められている。

1.3 新しい測定手法を適用した測度と同一用語による構成概念

従来型の心理尺度では、主観的な評価を尋ねる質問紙型の項目が主流である。また、学習の達成度などを測定するには、問いに対する正解をもとめるいわゆるテスト型の問題形式がとられることが多い。

しかし近年は、それらの方法とは異なる測定手法を適用したテストが作成されるようになってきた。一例として、潜在連合テスト (Implicit Association Test, IAT) が挙げられる。

このIATでは、測定対象とする構成概念に関係する属性語同士を分類する課題を課して、そこでの判断にかかる反応時間を指標として用いる。この方法を用いると、従来型の質問紙では生じてしまうとされる、社会的望ましさに影響を受ける回答の歪曲や抑制を、回避できるメリットがあるとされる。

藤井・上淵 (2010) は、暗黙の知能観の査定を目指し、個人の意識的な統制が及びにくいとされる潜在連合テスト (IAT) の手法を用いて、知能観 IAT の開発を試みている。この IAT では、知能観に関係すると思われる単語を用いている。具体的には、知能の属性語として、学力、知識、想像力、知恵、判断力を、運の属性語としては、まぐれ、運命、星回りを、増えるの属性語として、増加、成長などを、変わらないの属性語としては、固定、不変などといった単語を用いている。これらの属性語に対してグループ分け課題を課し、そこでの概念同士の組み合わせの違いの間で生じる反応時間の差から、概念間の連合強度を測定する。そして、その違いの様子から、当該の概念に係わる個人間の差異を測定しようとしている。分析では、知能観に係わる質問紙による顕在的な測度と、潜在的測度としての知能観 IAT での結果を比較し、IAT による測度の独自性を主張している。

相川・藤井 (2011) は、社会的心理特性であるシャイネスの測定を試みている。そこでは、先の藤井・上淵 (2010) と同じように、潜在連合テスト (IAT) を取り入れている。そして、シャイネスの測定のために、自己や他者、シャイか社交的かに関連する属性語を用意し、刺激語のグループ分け課題を行ない、概念同士の組み合わせの違いから生じる反応時間の差から、概念間の連合の強さを測定する日本語版シャイネス IAT を提案した。そして、質問紙などで測定されるシャイネスの顕在的側面は意識的な統制的処理を、一方、IAT で測られる潜在的側面は無意識的な非統制的処理を予測するものとしている。

この潜在連合テスト (IAT) などによる、新しい手法を用いた心理測度の提案は、挑戦的な取組みの新機軸として興味深い。しかしその一方、測定される構成概念そのものの存立が揺らいでしまう危うさも同時にはらんでいる。

すなわち上記の研究では、暗黙の知能観や、シャイネスといった単一の用語で規定した構成概念を対象として議論している。しかし、質問紙法による顕在的な測度と、潜在連合テスト (IAT) で測られる潜在的測度が示すものは性質が全く異なるという主張が強調さ

れるほど、単一であるはずの構成概念の二面性が際立ってしまう形となる。同一概念の量的な程度の差異を示す測度が、測定の方法の違いによって意味が異なるという状況には、やはり困惑せざるを得ない。異なるものに対して同一の呼び名を称するのは、科学的用語としても適切ではなかろう。

もし似て非なる概念を反映しているのだとすると、何を目途にその測度の妥当性の検証を行なえばよいのだろうか。少なくとも、構成概念と測定尺度間の関係の整理、必要に応じて構成概念の見直しを伴う再定義が必要となるのではないだろうか。

2. 測定尺度の妥当性検証の深化に向けた機運

本年の測定尺度に関連する特徴的な動向としては、尺度の妥当性の検証の必要性を訴えと共に、具体的な検証方法を提案するセミナーやシンポジウム、研究発表が行なわれている点を挙げることができよう。

2.1 心理尺度の氾濫と過剰供給への懸念

教育心理学会の第 53 回総会では、研究委員会企画によるチュートリアルセミナーが催された。そこでのテーマは、尺度の作成・使用と妥当性の検討であった (吉田・石井・南風原, 2011)。このテーマの設定の根底にある問題意識としては、心理学的な構成概念や測定尺度の過度の氾濫や過剰供給に対する危機感があったのではないかと忖度するところである。

1 章でも示したように、尺度作成研究の隆盛は並々ならぬものがある。南風原 (2011) はこの状況に際して、次のような指摘を紹介している。すなわち、心理尺度は需要を大幅に上回って供給され、バランスが崩れてしまっている。提案されただけで使用されない尺度のほうが多い。少数回しか使われない尺度が多数氾濫する状況は、科学として好ましくない。心理テストが次々に開発されてはいるが、妥当性が高く、使用に耐える尺度は少ない。また研究者は新しい構成概念について、疑わしく感じながらも、既存の構成概念とは異なっているのが当然とみなしがちである。などといった、尺度や構成概念の氾濫に対する危惧である。

一方で、測定尺度の評価の研究は、尺度作成の研究に比べて非常に少ないという問題を指摘している。このような状況に至っている背景として、開発研究の方が評価研究よりオリジナリティが高いと評価されがちな風潮、尺度作成者が提供する評価結果への過度の信頼があったとした。そしてその帰結として、ユーザが尺度を選択する際の情報が欠如し、クリティカルな評価に耐えない尺度が使われ続ける現状が生じているとし

た。

そこで南風原(2011)は、現在の状況を生じせしめている尺度作成のルーティン化された手続きを問い直し、尺度作成の本来の目標に照らして、より望ましい尺度作成・使用・検討のあり方を提案している。ここでは妥当性の捉えどころの無さからの脱却を企図し、尺度得点の成り立ちから信頼性と妥当性を整理している。

まず、尺度得点(x) = 測定したい構成概念を反映する成分(c) + 系統誤差(s) + 偶然誤差(e)の形に整理した。そして、尺度得点(x)全体の分散に占める成分(c)の分散の割合が妥当性にあたること。成分(c) + 系統誤差(s)で表される真値(t)の分散が、尺度得点(x)の分散に占める割合が信頼性であることを明示した上で説明を行なっている。

また、尺度得点(x)、測りたい成分(c)、系統誤差(s)、偶然誤差(e)をベクトルで表現することにより、相互の相関をベクトル間の角度として表し、直感的な理解を促している。この数量的な表現による土台をしっかりと固めた上で、偶然誤差(e)を減らすことによる信頼性の向上、系統誤差(s)の減少による妥当性の確保、という一貫した捉え方を提示している。

そしてまず、偶然誤差(e)の発生をローカルな要因とグローバルな要因に区別することで、 α 係数に代表される内的整合性に対応した信頼性と、再検査信頼性の違いを整理した。さらに、内的整合性の向上を図ると妥当性が低下してしまう帯域幅と忠実度のジレンマについても言及している。

次に、妥当性の低下に繋がる系統誤差の増大要因について、項目の内容や測定方法の偏り、反応バイアス、他の構成概念の反映などを挙げている。そして、他の構成概念の影響の除去に留意した弁別的妥当性の検証の必要性に言及している。

項目分析時の実際の手続き上の留意事項としては、特定の項目に天井効果・床効果などの反応の偏りがあっても、構成概念を反映する良い項目である可能性もあるので機械的な判断は避けるべきこと。項目得点の分布の歪みについては、因子分析に含めることの可否と構成概念の測度としての有用性の有無とは別問題であることなどを述べている。

また、尺度構成と因子分析の扱いについても触れている。そこでは、探索的因子分析自体の有用性は認めながらも、因子が構成概念と同一視されてしまうと、妥当性の観点からは問題が生じると指摘している。そして既存尺度の評価においては、階層因子分析や2次

因子分析といった確認的因子分析の適用が紹介された。

他の変数との相関に基づく妥当性の検証については、2つや3つの変数との相関の単純な比較だけでは不十分だとした。その上で、弁別の必要がある関連尺度の得点の影響を除いた上で、外部変数との相関(部分相関)などを求めるなど、妥当性への脅威に直接応える工夫によって、はじめて意味のある妥当性情報が得られるとしている。

尺度作成研究の論文の評価にあたっては、計画段階や作成段階から妥当化のためにどのようなステップを踏んだのか。妥当性の証拠として、focused & riskyな予測を、どれだけデータで検証できたのかに留意すべきであるとしている。

2.2 心理尺度の妥当性検証の現況

先の1章で紹介された尺度作成に係わる研究では、少なくとも学会誌に掲載されているものについては全て、尺度の信頼性の検討は元より、妥当性の検証に係わる記載がある。しかし、一言で妥当性と言っても、そこで論じられている妥当性の内容は、研究ごとにまちまちで多岐にわたっていた。また、その検証の手続きも多様である。

そこで検討されている妥当性の内容については、ざっと並べてみただけでも、依存性妥当性、交差妥当性、併存性妥当性、構成概念妥当性、内容的妥当性と実に様々であった。また、その検証のための手法としては、探索的因子分析や確認的因子分析、共分散分析、他の尺度の変数との単純な相関の比較、また専門家による評価といった手続きによるものであった。

しかしながら、南風原(2011)の言うところの予測的妥当性に該当するものは、ほとんど見られなかったように思われる。研究手続きの方法を読む限り、尺度作成のための調査時に、一度に収集できる資料内での分析検討に留まっているのが大多数であったと思われる。また、検討している妥当性の内容と、用いられている検証手続きの手法との間に、齟齬が感じられるものも見受けられた。

ところで高橋・山形・星野(2011)は、実施が容易であるという理由のみで大学生の自己回答を対象とするシンプルなサンプル収集の姿勢を、「大学生心理学」といささか揶揄的に呼んでいる。先の調査実施上の制約から生じていると考えられる妥当性検証の不全感は、その感触と同根かも知れない。そして、それは現在の心理学研究の状況の反映であるのかも知れない。

しかし、この妥当性検証の迷走とも呼ぶべき状況の原因は、研究者の側にだけあるのではなく、その内

容を評価する側の問題でもある。

2.3 妥当性の概念をめぐる同床異夢

研究論文の評価に関連して、教育心理学会の第53回総会では、教育心理学研究にもとめられる質の高い研究論文とは何か、というテーマで研究委員会企画のシンポジウムが行なわれている(服部・橋本・高橋・青山・大塚・浜谷・堀田, 2011)。

このシンポジウムの中で、大塚は、非常に哲学的で難解な論稿であるがと前置きした上で、尺度の妥当性の検証に取り組んで行くには、Messick (1989 池田訳 1992) による構成概念妥当性の捉え方と実証的な検証方法の考え方が、今後依拠していくべき方向性の1つを示しているとした。

このMessickの論稿は、『教育測定学』の中の第3章「妥当性」であり、翻訳も出版されているので尺度研究に携わる研究者は、まず実際に触れてみるのが肝要であろう。その中で、歴史的に多様だった妥当性概念の議論が整理され、構成概念妥当性の考え方に統合されてきた経緯を概観することにより、未だに錯綜している妥当性の用語や意味、関連性を見通すことが容易になると思われる。

この論稿の中には、構成概念妥当性を数量的に検証するアプローチの1つとして、多特性多方法行列(Multitrait-Multimethod Matrix) についての説明もある。それは、尺度の信頼性ととともに、収束的妥当性と

弁別的妥当性を同時に直感的にも評価して検討できる実証的なアプローチであると言える (Table 1)。

先の大塚は、そのような数量的な手続きをふまえた実証的な妥当性の検証方法の方向性を意図してMessickの論稿の紹介をしたものと推察される。一方、同じシンポジウムのパネリストの浜谷も、大塚と同じMessick (1989 池田訳 1992) の論稿を引用した講演を行なっている。その中で浜谷は、障がい児への教育支援の立場から、妥当性については、より結果的な側面に特に着目し、数理的な検証よりも、むしろ社会的役割の評価の重要性に関連した記述を主に引用した。

今回、妥当性の検証について、奇しくも同一の論稿に基づく提案がなされた訳であるが、シンポジウムのパネリストの間でさえ、妥当性概念の持つ意味合いが立場によって異なっている現状が垣間みられた瞬間であった。これはある種の同床異夢とでも言えるもので、研究論文の内容を評価する側にあっても、妥当性概念の扱いが一筋縄ではいかない現状を象徴しているとも考えられる。

2.4 妥当性検証の研究の進展と深化

前節にある構成概念妥当性の検証のための多特性多方法行列を用いた方法は、収束的妥当性や弁別的妥当性を統合的に評価できる利点がある。しかし、そのままでは、方法・特性の分散を定量的に評価できず、また方法・特性の共分散や交互作用も評価できない、と

Table 1 多特性多方法行列 (Multitrait-Multimethod Matrix) の仮想例 Messick (1989 池田訳 1992) より引用・補足

いった欠点がある(村山, 2006)。近年, これらの欠点を乗り越えるための研究も進められてきている。

久保(2011)は, 多特性多方法行列に対して確認的因子分析の枠組みを適用することによって, 信頼性, 収束的妥当性・弁別的妥当性を検討する方法を提案している。分析モデルにある種の制約を入れることにより, 観測変数の分散を特性因子・方法因子・誤差に分解し, 信頼性係数や妥当性を評価する係数をもとめている。

また一方, 構成概念の意義そのものの評価, という観点からの妥当性検証のニーズも指摘されている。

高橋・山形・星野(2011)は, パーソナリティ特性研究が, 経済学や疫学との協調の中でさらに発展し得る可能性について展望している。そこでは, 認知能力よりもパーソナリティ特性の方が, 労働・経済市場での社会的な成功に係わる結果変数を予測できるとする研究事例や, パーソナリティ特性が精神疾患や疾病の発症などの健康リスクの予防に貢献しうる知見について, 紹介している。そして, パーソナリティ特性はある程度変えられるかも知れないといった介入可能性から, 様々な予測や予防への媒介変数としてのパーソナリティ特性の役割が期待されていると述べている。その上で, このような他の学問分野との協調場面では, 予測力のある変数としてのパーソナリティ特性が必須であると共に, 広範な学問領域をカバーする観点では, 特性変数の予測的妥当性への注目度が特に高いとしている。

3. 教育評価・テストに係わる研究の進展

学校教育の現場で, 心理学的な測定・評価の研究と最も密接な関係があるのは, 教育評価の場面である。テスト理論は, 学力の適切な測定を目的とする試験を設計する上での科学的な基盤を与える。この教育評価やテスト理論に係わる研究は, その発展に伴って細分化が進んでいる。従来, 試験に係わる教育評価の研究は, 教育心理学会や心理学会などでも数多く発表が見受けられたが, 最近はテスト学会などにその発表の場が動きつつあるように思われる。

3.1 テストの技術面に関する研究

テスト理論や統計手法の開発, IT技術の活用など, テスト技術に軸足を置いて進めている研究を追ってみる。

登藤(2010)は, テストの形式として大問の中にいくつかの問題項目が含まれており, 項目応答理論の局所独立性の仮定が満たされない場合の, 潜在特性の推

定への影響について検討した。項目間の依存関係を許容する2つのモデルの下でシミュレーションを行ない, 潜在特性推定量へのバイアスや, 影響を与える要因について検討した。

野口・熊谷(2011)は, 共通受験者デザインを適用したテストの等化をする場合に頻繁に適用される Means & Sigma 法の改良を試みている。そこでは, 推定尺度値の分散に含まれる誤差成分の大きさを推定して取り除くという補正方法を提案し, シミュレーションによって評価し, その優位性が見られる条件を整理して示した。

大津・橋本(2011)は, 欠測値に対応できる非線形因子分析を開発している。その統計手法を適用し, 本来, 共通受験者の情報が存在しない大学入試センター試験の英語問題の, 異なる年度間の難易度比較を試みた。そこでは, 年度間で異なる試験の共通受験者の確保のため, モニター調査を別途実施した。その情報を用いて, 提案方法と IRT でそれぞれの試験の難易度の推定を行い, 精度の比較評価を行なっている。

さて, 入学試験の得点と入学後の成績の相関の値は, 入試で不合格になった者は欠測値となるために, しばしば想定よりも小さい値となることが多い。岡田・繁樹(2010)は, そのように過小評価される相関係数の真の値を推定する方法として提案されている, 最尤推定法とベイズ推定法の2つの方法について検討した。特に小標本の場合での挙動について, 数値的なシミュレーションにより比較して評価を行なっている。

孫・高橋・柿沼(2011)は, 教科内容を理解するために必要となる個々の認知的技能(attribute)の測定を目的とした, ルールスペース法に基づく認知診断テストを提案している。このテストでは, 教師や認知心理学の専門家が抽出した attribute を用いて, 問題×attribute 行列形式のリストを用意する。そして, 問題の正誤情報から, 受験者ごとの attribute の習得状況を同定する。この研究では, TIMSS の試験データにこの分析方法の適用を試みている。

寺澤・吉田・三宅・岩本・池田(2011)は, マイクロステップ計測法と呼ばれるドリル学習を中心に据えたテスト技術とITの活用によって, 子ども一人一人の学力や感情状態を連続的に追跡して測定するシステムの提案を行なった。吉田(2011)は, 英単語の学習にマイクロステップ計測法を適用し, その継続的なデータの蓄積から, 形成的評価にあたっての利点を示した。三宅・寺澤(2011)は, 国語の漢字の読みにそれを適用しながら, その長期継続した実践の中で自己効力感と

の関連も検討した。岩本・寺澤 (2011) は、不登校生徒の在宅学習支援にマイクロステップ計測法を導入し、Web と紙媒体によるハイブリッドなやり取りを通じて、長期にわたる継続性の利点を述べている。寺澤 (2011) は、これらの実践活動を通じて、学術研究と実践が融合した、大学、小・中学校、家庭を繋ぐ新たなインフラの構築を提案している。

山田・村井・杉澤・寺尾・高橋・池田 (2011) は、本稿で扱っている測定・評価領域の基盤となる心理統計の教育場面での Web の活用に向けた企画発表を行っている。寺尾は、統計教育での携帯端末の活用事例として、インタラクティブな授業の実現、セカンドモニタとしての利用、Ustream による授業配信などを紹介している。高橋は、信州大学の心理統計の授業で e-Learning システム (Moodle) を利用している事例を示し、授業の各種資料の提供、コミュニケーションの場としての活用例を示している。杉澤は、かねてより大学で心理統計の授業を担当する教員が講義のために作成したテスト問題を収集して、相互に共有して利用するための web データベースを作成してきた。今般、統計指標の情報付与の廃止、検索機能の拡充、ユーザによるコメント記入機能の追加などの改良点をまとめ、運用事例などと共に紹介している。

3.2 テスト理論の進展と現実面での制約

テスト理論などの技術面の研究が進められる一方で、現実の社会では、社会的な責任を担いつつ試験を運営実施し、試験問題の品質の管理保守を行なっていく実務的な立場もある。実際の試験の運営場面では、異なるテスト間の成績を比較可能にするための等化の作業の中で、テスト理論が前提としてもとめる条件を満たすのが手続き的にも経費的にも難しいことが多い。ここでは、自由な発想の下で進められる理論的な研究とは異なり、即時的に実現可能な対応が求められる現実的な制約がある。

第9回のテスト学会では、このテスト等化の理論と実践についての企画セッションが設けられた (前田・加藤・野澤・野上・佐藤・野口, 2011)。野澤 (2011) は、等化における統計指標について、アメリカの主に中西部での影響力が大きいテスト実施組織である ACT の研究者として、最近の理論面での解説を行なった。

一方、野上 (2011) は、日本で開発されて運営されているコンピュータ適応型の英語テスト CASEC を巡る状況を説明している。特に項目バンクの維持にかかる課題、すなわち、新聞のためのデータ収集における受験者確保の難しさ、限られた項目数の中でのアン

カー項目の配置の限界、項目バンクの維持に係わるコスト面での制約など、現実には直面する課題を提示した。

また佐藤 (2011) は、日弁連法務研究財団が主催する法科大学院全国統一試験における事例を紹介した。この試験は、年内に2回実施されるものであり、第1回と第2回の得点の等化がもとめられる。その中で受験者数の状況や試験問題の構成など、この試験に特有な各種条件の制約などの検討を通じて、下位テスト得点ごとの等パーセンタイル法による等化を採用するに至った経緯を紹介している。

3.3 テストの社会的役割への期待

教育テストや試験に対して社会が期待する役割は、世界的なグローバル化の潮流の中で、国際的観点からの比較がもとめられる部分が大きくなっている。先のテスト学会では、「世界的に見た教育テストの今とこれから」と称する企画シンポジウムが開催されている (渡部・Yamamoto・李・川嶋, 2011)。

アメリカの最も有力な試験実施機関の1つである ETS の Yamamoto (2011) は、国際的で大規模な能力評価 (Large Scale Assessment) に関連する試験の数と頻度の増大を指摘している。そして数学や理科の学力に関する国際的な比較研究であるところの TIMSS、言語の読み能力を対象とする PIRLS をはじめ、生徒学生の国際的な評価プログラムである PISA、成人の能力査定を目的とする PIAAC などについて、その実施団体や役割、歴史や特徴を紹介している。

李 (2011) は、韓国の大学入試センター試験にあたる大学修学能力試験について紹介した。そこでは、大学修学能力試験の歴史的経緯、改訂の背景と方向、水準別試験の提供、科目の変更、複数回実施に向けた検討状況などについて整理して示している。

川嶋 (2011) は、大衆化が進んできた日本の大学における学士の水準の維持のために、高等教育でのテストの活用を論じている。そこでは、各授業科目におけるアセスメントを、レベルの目安を数段階に分けて記述した達成度の判断基準であるルーブリックを活用して厳格に行なう方向と、国際的に標準化されたテストを導入して、大学生の学習の達成度を測定しようとする方向の、2つの方向性が考えられるとしている。

3.4 試験・テストの教育・社会への波及効果

試験やテストは、本来、学習成果や能力などを正確に測定することが目的である。しかし、その試験の存在やあり方そのものが、教育内容は元より、学習活動そのものにも影響を与える。さらに、受験者の処遇を左右するハイスタークスの試験であるほど、社会的な

インパクトを持つ波及効果が生ずる。

Messick (1996) は、妥当性の検証の重要性と共に、教育アセスメント、試験、テストが学校や会社で実際に運用される場合には、その教育や社会に与える波及効果 (washback) をトレースする必要性を指摘している。そして、良いテストの必須条件として、良い方向の波及効果をもたらすことを見極めることが重要であるとしている。

翻って、教育評価やテストに関連したわが国の研究をさらに概観してみる。

鈴木 (2011) は、テストの評価基準や評価目的を学習者に示すことで、動機づけやテスト観の変容が見られ、より良い効果が生ずるかを検討した。中学生に数学のテストの結果を返却する際に、ループリックを示すことによって、テスト観の変容が促され、理解への指向性が進んで高成績を上げることができたとしている。

さて、わが国では、試験やテストが行われた場合、全ての問題が開示されることは、社会的な常識として深く浸透している。このことは、学力の年次推移を検討するような場合、同一の問題項目を使えないために、その比較には大きな困難をもたらす。石井・安永 (2011) は、そのようなわが国のテスト文化のもとで、学力の経年比較を実現する可能性を検討した。そこでは、全国テストと自治体テストをうまく組み合わせることで、テストのリンキングが可能なデータ収集のデザインを提案している。

萩原 (2010) は、全国的に実施された小学生を対象とした、英語の聞き取りテストに関するデータを分析している。そのテストは、小学校での外国語活動が必修化される前に実施されたものであった。そこでは、各学校での英語活動の取組みに、導入済み・未導入の違いが想定されるため、学校間の生徒の英語能力をその試験得点だけで判断するのは適切でない。そこで、項目困難度に学校レベルでの散らばりを仮定した項目反応モデルを適用した方法の提案を試みている。

内田・杉澤・伊藤 (2010) は、大学入試センター試験の英語リスニングテストが測定する学力を検討している。そこでは、2006 年から新たに導入された英語リスニングと既存の教科科目の学力の関係性について主成分分析を行なって図示すると共に、その年度推移を追跡している。また、外国語を使って何がどの程度できるのかという行動のレベルを尋ねる調査を行ない、読解行動の熟達度と聴解行動の熟達度の指標を収集し、英語筆記試験とリスニングの試験成績との相互の関係

を示した。その結果は、読む能力、聴く能力の構成概念に関する多方法多特性行列の形を成しており、それによりリスニングテストの妥当性についても評価を行っている。

倉元 (2011) は、教育政策と学力測定の技術について多面的な視点から議論している。そこでは、全国学力・学習状況調査、大学入学試験、医療職業人養成課程の統一テスト、法科大学院入学試験などの試験について、目的や歴史的経緯をまとめている。その中で、政治的な企図、イデオロギーや理念、社会的な要請・制約、世論の動向、公平性などの社会通念、歴史的な経緯などの複眼的な観点から、学力測定に関わる試験を論じている。

倉元はその中で、「全国学力・学習状況調査については、同一冊子による一斉同時実施方式しか念頭になく、その他の可能性について最初から意図せず除外してしまっている」という指摘を紹介している。この指摘にある制約のため、この学力調査の設計思想は、学力の経年比較への道筋を閉ざしてしまっている。この行政主体の調査における教育評価測定の専門家の不在、この現状を批判することは簡単である。しかし、真に必要なことは、教育評価測定の研究が、教育心理学会の領域に留まらず、より広い社会でのプレゼンスを高めることであると自戒する必要がある。

さてこのように、様々な制約を孕む社会で、実際に運用されているテストや試験の場面で発生している具体的な問題に向き合い、その解決を目的とした研究が行なわれつつある。これらの研究は、理論的な研究の進展と共に、教育心理学そのものの課題でもあるところの、研究と教育実践の社会的な繋がりを考える上で大切な方向性の 1 つを示しているものと考えられる。

引用文献

- 安達智子 (2010). キャリア探索尺度の再検討 心理学研究, 81, 132-139.
- 相川 充・藤井 勉 (2011). 潜在連合テスト (IAT) を用いた潜在的シャイネス測定の試み 心理学研究, 82, 41-48.
- 赤田太郎 (2010). 保育士ストレス評定尺度の作成と信頼性・妥当性の検討 心理学研究, 81, 158-166.
- 藤井 勉・上淵 寿 (2010). 潜在連合テストを用いた暗黙の知能観の査定と信頼性・妥当性の検討 教育心理学研究, 58, 263-274.
- 南風原朝和 (2011). 尺度の作成・使用と妥当性の検討 日本教育心理学会第 53 回総会 研究委員会企画

- チュートリアルセミナー発表資料
- 萩原康仁 (2010). 学校での多様な取組と年間授業時間数を考慮した小学校の英語活動におけるテストデータの分析 日本テスト学会誌, **6**, 75-86.
- 服部 環・橋本創一・高橋知音・青山征彦・大塚雄作・浜谷直人・堀田香織 (2011). 研究委員会企画シンポジウム3: 教育心理学研究に求められる質の高い研究論文とは何か—『教育心理学研究』の現状と課題について— 教心 **53**, 6-7.
- 石井秀宗・安永和央 (2011). 全項目が開示されるテスト文化のもとでの得点分布の経年比較 日本テスト学会誌, **7**, 23-35.
- 井潤知美 (2010). Parenting Scale 日本語版の作成および因子構造の検討 心理学研究, **81**, 446-452.
- 岩本信喜・梅津亜耶子 (2011). Theory of Intelligence Scale 日本語版の作成 教心 **53**, 535.
- 岩本真弓・寺澤孝文 (2011). 在宅学習支援を入口とした不登校生徒への支援 日本テスト学会第9回大会発表論文抄録集, 32-35.
- 川嶋太津夫 (2011). 高等教育における「テスト」活用に関する国内外の動向—質保証の高まりを背景として— 日本テスト学会第9回大会発表論文抄録集, 48-49.
- Yamamoto, K. (2011). Brief History, methodology, current states and future directions of international large scale assessment 日本テスト学会第9回大会発表論文抄録集, 42-43.
- 小浜 駿 (2010). 先延ばし意識特性尺度の作成と信頼性および妥当性の検討 教育心理学研究, **58**, 325-337.
- 近藤美鈴・金子智栄子 (2011). 母親の育児対処能力に関する研究 I—SOC 測定尺度の構成について— 教心 **53**, 100.
- 今野紀子・土肥紳一・宮川 治 (2011). 数学学習におけるモチベーション評価尺度作成の試み 教心 **53**, 541.
- 高坂康雅 (2011). 共同体感覚尺度の作成 教育心理学研究, **59**, 88-99.
- 久保沙織 (2011). 多特性多方法行列を利用した確認的因子分析モデルによる信頼性および収束の妥当性・弁別的妥当性の検討—方法因子の和が0になるという制約のもとで— 日本心理学会第75回大会発表論文集, L(8).
- 倉元直樹 (2011). 教育政策と学力測定 of 技術 日本児童研究所・平木典子・稲垣佳世子・斉藤こずゑ・高橋恵子・氏家達夫・湯川良三 (編) 児童心理学の進歩 Vol.50 [2011 年版] (pp. 200-217) 金子書房
- 前田有希子・加藤健太郎・野澤雄樹・野上康子・佐藤喜一・野口裕之 (2011). テスト等化の理論と実践 日本テスト学会第9回大会発表論文抄録集, 11.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). (pp. 13-103). New York: American Council on Education and Macmillan. (メシック, S. 池田 央 (訳) (1992). 妥当性 池田 央他 (監訳) 教育測定学上巻 (pp. 19-145) みくに出版)
- Messick, S. (1996). Validity and washback in language testing. Educational Testing Service Research Report, RR-96-17, 1-18.
- 三宅貴久子・寺澤孝文 (2011). 評価と一体化した真の指導・支援をめざして 日本テスト学会第9回大会発表論文抄録集, 30-31.
- 村山 航 (2006). 妥当性概念の展開 日本テスト学会第7回研究会報告資料 <<http://www.jartest.jp/study7.html>>
- 並川 努 (2011). 継時的比較の個人差—継時的比較志向性尺度の作成と検討— 心理学研究, **81**, 593-601.
- 野上康子 (2011). 等化を前提とした試験運用における項目バンクの維持に関する諸問題 日本テスト学会第9回大会研究発表資料
- 野口裕之・熊谷龍一 (2011). 共通受験者デザインにおける Mean & Sigma 法による等化係数推定値の補正 日本テスト学会誌, **7**, 15-22.
- 野澤雄樹 (2011). 等化の実践における統計指標の利用について—等化デザインおよび等化方法との関連に焦点を当てて— 日本テスト学会第9回大会発表論文抄録集, 62-63.
- 岡田謙介・繁榊算男 (2010). 小標本における選抜効果を補正する相関係数の推定について—最尤推定法とベイズ推定法のシミュレーションによる比較— 日本テスト学会誌, **6**, 63-74.
- 小山内秀和・岡田 斉 (2011). 物語理解に伴う主観的体験を測定する尺度 (LRQ-J) の作成 心理学研究, **82**, 167-174.
- 大津起夫・橋本貴充 (2011). 非線形因子分析によるセンター試験英語問題の難易度比較 日本テスト学会誌, **7**, 1-14.
- 李 庸伯 (2011). 大学修学能力試験の現在と未来 日本テスト学会第9回大会発表論文抄録集, 44-47.

- 崔 玉芬・庄司一子 (2011). 中学生用自己防衛 (Self-Regulation) 尺度の検討 教心 **53**, 534.
- 佐藤喜一 (2011). 法科大学院全国統一適性試験の年度内等化について 日本テスト学会第9回大会発表論文抄録集, 64-65.
- 孫 媛・高橋智子・柿沼澄男 (2011). 「目標に準拠した評価」のための認知診断テストの開発 教心 **53**, 542.
- 鋤柄増根 (2011). シミュレーションデータによる心理検査の短縮版と完全版の等価性の検証方法 日本心理学会第75回大会発表論文集, 482.
- 鈴木雅之 (2011). ルーブリックの提示による評価基準・評価目的の教示が学習者に及ぼす影響—テスト観・動機づけ・学習方略に着目して— 教育心理学研究, **59**, 131-143.
- 高橋雄介・山形伸二・星野崇宏 (2011). パーソナリティ特性研究の新展開と経済学・疫学など他領域への貢献の可能性 心理学研究, **82**, 63-76.
- 谷 伊織・並川 努・脇田貴文・中根 愛・野口裕之 (2011). 日本語版 State-Trait Anxiety Inventory for Children (STAI-C) における特性不安尺度短縮版の作成(1)—IRTを適用した短縮版の作成— 教心 **53**, 533.
- 寺澤孝文 (2011). 学術研究と実践が融合した新たなインフラの構築 日本テスト学会第9回大会発表論文抄録集, 36-39.
- 寺澤孝文・吉田哲也・三宅貴久子・岩本真弓・池田央 (2011). 企画シンポジウム1：子どもを見守る動的テスト技術—子ども一人一人の学力と感情状態の連続測定— 日本テスト学会第9回大会発表論文抄録集, 8.
- 登藤直弥 (2010). 局所独立性の仮定が満たされない場合の潜在特性推定への影響 日本テスト学会誌, **6**, 17-28.
- 内田照久・杉澤武俊・伊藤 圭 (2010). 大学入試センター試験のリスニングテストが測定する学力の布置 日本テスト学会誌, **6**, 103-111.
- 渡部 洋・Kentaro Yamamoto・李 庸伯・川嶋太津夫 (2011). 企画シンポジウム2：世界的に見た教育テストの今とこれから 日本テスト学会第9回大会発表論文抄録集, 9.
- 山田剛史・村井潤一郎・杉澤武俊・寺尾 敦・高橋知音・池田 央 (2011). 心理統計教育におけるウェブの活用 教心 **53**, 642-643.
- 吉田哲也・寺澤孝文 (2011). 1000語の英単語を覚えるためには何ヶ月必要か? 日本テスト学会第9回大会発表論文抄録集, 26-29.
- 吉田寿夫・石井秀宗・南風原朝和 (2011). 研究委員会企画チュートリアルセミナー：尺度の作成・使用と妥当性の検討 教心 **53**, (28).

付 記

本稿で紹介した一部の学会発表について筆者が直接捕捉できなかった発表内容に関しては、京都学園大学の行廣隆次先生から貴重な情報を提供していただきました。ここに深く感謝いたします。