首都大学東京人文学報
2016, No. 512-4 15－26

原　著

妥当性理論の歴史的変遷と心理学研究への適用に関する一考察 －*Standards* を中心に－

A view on the history of validity theory and its application to our psychological research: around *Standards*

平　井　洋　子

1．はじめに

心理測定において，測定の妥当性は根幹的で最も重要な性質である（Angoff, 1988）。テスト[1]得点が実際に反映する内容が想定と食い違えば，テスト得点に基づく正しい判断は成り立たない。教育・心理の領域における測定手続きの規準を定めた *Standards for Educational and Psychological Testing*（AERA, APA, & NCME; 1985, 1999, 2014 など，以下 *Standards* と略す。）においても，テストの開発において妥当性への配慮は最も根本的な事項であると明記され（例えば AERA, APA, NCME, 2014, p.11），妥当性に関する章は基本的に冒頭に配置されている。

その測定の妥当性についての議論が，近年活発である。妥当性を中心的テーマに据えた書籍といえば，以前は Wainer & Braun (1988) ぐらいであった。しかし今世紀に入ると，目についたものだけでも，Braun, Jackson, & Wiley (2002), Borsboom (2005), Lissitz (2009) と相次いで発行され，ここ 3 年間では，Chatterji (2013), Markus & Borsboom (2013), Newton & Shaw (2014), Zumbo & Chan (2014) と，かつてない密度で刊行されている。雑誌論文の方も同様で，表 1 に示すように，妥当性に関する特集号が次々と発行されている。これらの議論の方向は，現在主流となっている妥当性理論に対して批判的なものから，主流となっている理論を現場でどのように応用すべきかまで，多岐に渡る。

Standards は，心理検査や教育測定に密接に関わる 3 学会（アメリカ教育学会 AERA，アメリカ心理学会 APA，全米教育測定評議会 NCME）が共同で策定したものである。「テストの開発及び，テストとその実施を評価するさいの規準を定め，また得点解釈の妥当性を査定するさいのガイドラインを定める（*Standards*, 2014, p.1）」ことを目的としている。テストを開発する専門家やテスト出版者だけでなく，スポンサーや一般のテストユーザーにも，*Standards* を適切に満たすことが求められる。ただし *Standards* に記載される個々の規準は一般的な表現に留まり，具体的な手続きや数値をレシピのように示してはいない。アメリカ合衆国内の学会が定めたものであるが，いずれの学会も世界中の研究者が加入している大規模な学会であることから，実質的には世界標準と位置づけることができる。

Standards の歴史は古く，APA が発行した *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA, AERA, & NCME, 1954) と AERA と

表 1：近年発行された妥当性に関する特集号

雑誌名・巻号	テーマ（フォーカス論文のタイトルなど）
Psychological Assessment, 2005, 17(4)	Construct validity of psychological tests: 50 years after Cronbach and Meehl (1955).
Educational Researcher, 2007, 36(8)	Questions about the current unified theory of test validity.
Measurement: Interdisciplinary Research and Perspectives, 2008, 6(1-2)	Conceptual foundations of psychological measurement.
Measurement: Interdisciplinary Research and Perspectives, 2012, 10(1-2)	Clarifying the onsensus definition of validity.
Journal of Educational Measurement, 2013, 50(1)	Validating the interpretations and uses of test scores.

NCMEが共同で発行した *Technical Recommendations for Achievement Tests* (AERA & NCME, 1955) を前身として，1966年には3学会共同で *Standards for Educational and Psychological Tests and Manuals* が出版され，以降1974年，1985年，1999年，2014年と改訂されてきた。*Standards* が興味深いのは，その時代における「測定の品質に対する理論的潮流」と「テスト使用の現場」をうかがい知ることができる点にある。

近年の妥当性理論に関する活発な議論は，現在主流となっている妥当性理論のある意味での行き詰まりと，この時期が1999年版の *Standards* の改訂作業期間[2]にあたっていたことに無関係ではないだろう。現在主流となっている妥当性の考え方は，1980年代にS. Messickがそれまでの議論をまとめ，発展させたものが土台となっている。彼の妥当性理論を含むこれまでの妥当性理論の変遷については後に詳しく述べるが，現在主流となっている妥当性の考え方の要点として，ここでは，(a) 妥当性は内容妥当性，基準関連妥当性，構成概念妥当性などに分かれるものでなく，統一的なものであること，(b) テストに固有の妥当性があるのではなく，得点の解釈の仕方ごとに妥当性を考えるべきであること[3]，(c) 妥当性は程度として評価されるもので，有・無で議論されるものではないこと，(d) 妥当性の評価対象は，テストを使用したことで生じた影響まで含むこと，を指摘しておく（Cizek, 2012）。

これに対し近年の議論では，テストの使用やその結果まで含めると妥当性の概念が広がりすぎるという主張（Cizek, 2012）や，定義が曖昧なのでより明確化すべきだという指摘（Newton, 2012），測定される特性は仮説的な構成概念ではなく実在するとして，認識論的に異なる立場から測定を考える立場（Borsboom, Mellenbergh, & van Heerden, 2004），など，別な定義や概念化を探る議論が出されている。一方で，テストが受験者の特徴の把握という役割を超えて二次的な用途に使われている実態（例えば学力試験を教育プログラムの評価や教員評価に用いる場合）や，副次的であるが可能性の高い帰結（例えば教師が共通学力テストの対策に力を入れた結果，教える内容が限定された場合）があることを踏まえ，テスト使用の結果生じた事態を妥当性の評価に含めるのは当然という立場をとる主張もある（例えばBennett, 2012; Lane, 2012）。

いずれにせよ，これらの議論はほとんど日本に浸透せず，我が国の心理学研究の現場では，「内容妥当性」「基準関連妥当性」「構成概念妥当性」に分かれた妥当性概念が依然として使われ，妥当性検証の作業も，何をどこまですべきかわからず漫然と行われている観がある。

その原因として考えられるのは，妥当性の定義や議論が抽象的で時に哲学的なため難解であること，論文や *Standards* には具体的で詳細な事例がほとんど書かれていないため参考にしにくいこと，アメリカの教育・心理テスト事情を土台にして展開されているためイメージしにくいこと，などの要因があると考えられる。翻訳書（赤木・池田, 1993；メシック, 1992）や新しい妥当性理論を紹介する文献（平井, 2006；村上, 2003；村山, 2012など）が現れても，それを自分の研究にどう生かせばいいのか具体的にわからないために，仕方なく先行研究のやり方を真似ているというのが実態ではないだろうか。

本論文では，はじめに妥当性理論の歴史的経緯を概観し，そののち我が国の心理学研究の実態に合った妥当性の考え方を探る。日本の心理学研究において測定の妥当性が問題となる状況はさまざまであるが，最もよくある場合として研究者による心理尺度の作成と使用を想定して妥当性概念の適用を考える。

2．妥当性概念の歴史的変遷

妥当性概念の歴史的経緯については，Angoff (1988), Fiske (2002), Kane (2006), Newton (2012), Newton & Shaw (2014), Sireci (2009) などが詳しい。ここでは，これらの文献を参考にしつつ，その時代のテスト使用の実態にも注目しながら，妥当性がどのように概念化されてきたかについて概観する。

2－1．妥当性の基準関連モデル

20世紀の初頭，ピアソンが相関係数の公式を発表し，相関係数は画期的な統計手法としてテストの計量的研究に取り入れられた (Sireci, 2009)。この時代のテストは，将来のパフォーマンスを予測するためのものが多かった。ビネー・シモン尺度や陸軍の入隊検査をはじめとする，各種の採用・適性試験などである。入学や入隊した後のパフォーマンスを基準変数，テストを予測変数とし，その間の相関係数や回帰分析がテストの妥当性評価に用いられた。1920年代までには「テストは相関の高いものなら何に対しても妥当性がある」という考え方が成立し，1950年代あたりまでは，このような妥当性の考え方（"妥当性の基準関連モデル" と呼ばれることがある）が一般的であった (Sireci, 2009)。この考え方の好例として，ギルフォードによる妥当性の定義をあげておこう。

"a test is valid for anything with which it correlates. (Guilford, 1946, p.429)"

妥当性の基準関連モデルは，適切な基準変数が入手できれば，簡単で客観的かつ量的に妥当性が評価できる。しかし一方で，(a) 適切な基準変数の入手が必ずしも容易でない（Kane, 2006），(b) 基準変数の妥当性や信頼性が何も問われていない（Kane, 2006），(c) 基準変数との相関が高ければ，テストの内容が何であれ妥当性が高いとされる，などの問題があった。(a) は，例えば，アチーブメントテストではそれ以上に優れた基準変数が考えにくいとか，知能検査では基準を十分に概念化すること自体が困難であるという例を考えるとわかりやすい。(b) は，基準変数の妥当性を評価するには，それに対する基準変数を別に設定し，その基準変数の妥当性を保証するためにはさらに別の基準変数を導入して・・・となり，無限に連鎖してしまうという点である。また，実際に空軍パイロット養成の適性検査で基準変数の信頼性を調べたところ，ほとんどゼロだったという報告もある（Jenkins, 1946）。(c) は，テスト内容が突飛であっても相関が高ければよいという事態が防げないという問題である。

2－2．妥当性の内容モデル

基準関連モデルへの批判を受け，テストが測ろうとしている特性や，基準変数に反映される内容に対して，より関心が向けられるようになった。測定される特性の操作的定義やテストの内容分析などが行われるようになり（Sireci, 2009），いわゆる内容妥当性の考え方（"妥当性の内容モデル" と呼ばれることがある）が生まれたのである。

P. J. Rulon は，内容妥当性の初期の提唱者のひとりである。彼は定められた教育目標が存在するアチーブメントテストについて論じる中で，「テストの素材や（解答）プロセスが教育目標に記されているものと等しければ，そのテストにはそれ以上の妥当性の証明は必要ない（Rulon, 1946, p.291，カッコ内は著者追加）」としている。だたし Rulon は内容妥当性がすべてだと主張したわけではなく，テストの目的に応じて異なるエビデンスが必要になるという立場であった。そして「テストの妥当性は用途に応じて高くも低くもなりうる。妥当性の問題は，つまるところテストによって我々がしようとしていることをテストが行っているかどうかである（Rulon, 1946, p.290）。」とも述べている。この当時，用途に応じて妥当性の評価が変わりうること，よって異なるエビデンスが必要になることを指摘している点は，現在主流となっている妥当性の考え方に通ずるものであり，その先駆的なものとして注目される。

妥当性の内容モデルにも限界や批判がある。(a) 測定したい内容のドメインは通常大きく，テストにはその一部しか含められない。すなわちドメイン全体を過不足なくカバーするテストは多くない (Angoff, 1988)，(b) パーソナリティ特性などでは，ドメインの境界を明確に線引きすることが難しい (Angoff, 1988)，(c) 妥当性の評価が人による主観的評価で行われる。テスト開発者が行えば確証バイアスが入り込む (Kane, 2006)，(d) ドメイン代表性の根拠にはなるが，得点の意味を直接保証するものではない（Kane, 2006），などである。内容妥当性は，実際のテスト得点がどのようにふるまうかに関する実証的なエビデンスではない。表面的にはある特性を測っているように見えても，他の要因（方法因子や回答の構えなど）がどの程度入り込んでいるかについては何も語らない。Rulon が述べたアチーブメントテストのように，測定されるものが境界のはっきりした具体的な内容の集合であり，テストがその全域に渡る代表サンプルとなっているときには有効な考え方である。しかしそれ以外のときは，内容妥当性のエビデンス単独で妥当性を主張するのは難しいといえよう。

2－3．妥当性の構成概念モデル

1950 年代の初めは，基準関連モデルが成熟し，基準変数の内容に関する妥当性には内容モデルが用いられた時代であった。またアチーブメントテストなど，具体的な測定内容が線引きできる場合には，内容モデルが用いられた（Kane, 2006）。しかし L. Cronbach や P. E. Meehl は，テストに適切な外的基準が存在せず，かつ測定される内容のドメインが具体的に線引きできない場合があるとして，例えば投影法テストの解釈の妥当性はどう検証すればいいのかを取り上げた（Cronbach & Meehl, 1955）。パーソナリティ検査には明らかな外的基準がなく，サンプリングすべき内容のドメインも明確ではない。そこにあるのは測定したい特性をスケッチした理論のみである。測りたい特性は理論的に構成された概念である。このような場合，測定の妥当性をどう扱えばいいのであろうか。

構成概念妥当性という用語が初めて用いられたのは，*Standards* の前身のひとつ *Technical Recommendations for Psychological Tests and Diagnostic Techniques*

(APA, AERA, & NCME, 1954) である（Sireci, 2009）。この *Technical Recommendations* では，それまでの妥当性の概念や用語を抜本的に変え，妥当性を 4 つの「タイプ types」もしくは「属性 attributes」に分類して示した。すなわち，予測的妥当性，並存的妥当性，内容妥当性，そして構成概念妥当性である。このときの検討メンバーの中心的存在であった Cronbach と Meehl が構成概念妥当性の考え方をより敷衍した論文が，その後の妥当性理論に大きな影響を与えることになった Cronbach & Meehl (1955) である。彼らは「構成概念とは，テストにおける個人のふるまいに反映される，仮定された属性 postulated attribute である（p.283）」とした上で，

> 「構成概念妥当性の検証は，"操作的に定義されない" 属性や質の測度としてテストが解釈されるときは，常に行われなければならない。」
> 「テストにおけるパフォーマンスがどのような心理学的構成概念によって説明されるかを定める作業は，ほとんどすべてのテストにとって望ましい（Cronbach & Meehl, 1955, p.282）。」

と述べ，基準関連妥当性や内容妥当性のアプローチが難しい場合として，構成概念妥当性を提唱した。彼らは基本的には妥当性の "住み分け" を意図していたと思われるが，控え目な言い回しながら，テスト得点を構成概念に関連して解釈するときには，どのテストであろうと構成概念妥当性の検証が望ましいとも述べている。

では構成概念妥当性の検証はどのように行うのであろうか。測定したい特性が操作的に定義されない場合，それを取り巻く理論によってそのあり方が定位され，内在的に意味が決まる。その様子を定式化したものが，「法則論ネットワーク nomological network (Cronbach & Meehl, 1955)」である。法則論ネットワークは，(a) 観測可能なものどうしの関係，(b) 構成概念と観測可能なものとの関係，(c) 構成概念どうしの関係，の 3 種類の法則群から構成される。(c) は，関心下の構成概念を取り巻く理論に当たる部分と考えられる。法則は，統計的に表されていても，決定論的に表されていてもよい。関心下の構成概念は，ネットワーク内で何らかの観測可能なものに結びついていなければならない（(c) を通じて他の構成概念の指標に間接的に結びついているのでもよい）。構成概念妥当性の検証は，構成概念を取り巻く理論ごとエビデンスによって実証していくことである。構成概念を取り巻く理論が適切であれば，観測可能な指標の値は理論に基づく予測通りのふるまいをする。もしデータが予測に反したら，理論のどこかに不備があるか，指標の設定に誤りがあるか，補助的な仮定が犯されているか，のいずれかになる。つまり理論による予測が可能になるためにはある程度成熟した理論が必要となるが，テスト開発の段階では理論が未成熟なことが多い。実際 Cronbach と Meehl も，

> 「不安とは何か」は，それに関するすべての法則がわかったときにいえるであろう。我々はその法則を明らかにしている途中なので，正確に「不安とは何か」をまだいうことができない（Cronbach & Meehl, 1955, p.294）。

と述べ，法則論ネットワークが少しずつ "育って" いきながら，理論も測定も洗練されていくプロセスを考えている。

Cronbach & Meehl (1955) がその後に与えた影響は大きい。構成概念妥当性の概念が強く打ち出された点はもちろんであるが，ほかに，妥当性は 1 つの係数（たとえば妥当性係数や相関係数）で示されるのではなく，多くの実証的エビデンスのあり方によって示されるとしている点（Fiske, 2002）も重要である。彼らは，構成概念妥当性の検証は科学研究で一般に行われているアプローチと基本的に同じだと主張した。この考え方は，現在の *Standards* や妥当性検証のアプローチ（例えば Kane, 1992, 2009 など）にも広く取り入れられており，Cronbach & Meehl (1955) はその重要な土台のひとつと考えることができる。

Cronbach & Meehl (1955) が発表されると，妥当性理論の研究者たちから，構成概念妥当性はすべての心理・教育テストに該当するという声が出されるようになった（Sireci, 2009）。中でも有名なのは，J. Loevinger による

> 「予測的妥当性，並存的妥当性，内容妥当性は，いずれも根本的に場面依存的 *ad hoc* であるため，科学的見地からすると構成概念妥当性が妥当性のすべてである（Loevinger, 1967, p.78）」

という主張であろう。ここで "場面依存的 *ad hoc*" とは，例えば予測的妥当性や並存的妥当性は，基準変数に何を使用するかに依存し，内容妥当性は，それを評

価する人に依存するというように，個々の条件に左右されるという意味である。

もっともこうした構成概念妥当性を基本とする主張はすぐには主流とならず，1970年代半ばまでは，妥当性は予測的妥当性，並存的妥当性，内容妥当性，構成概念妥当性に分かれるものとして扱われていた。そのことを示すため，1966年版の *Standards* (APA, AERA, & NCME, 1966) から，基準関連妥当性，内容妥当性，構成概念妥当性の定義に近い記述を抜粋する[4]。

> 「内容妥当性は，教室での状況や教材について何らかの結論を下すとき，テストがどの程度それらを標本抽出しているかによって示される。」
> 「基準関連妥当性は，問題となる特性や行動の直接的な測度として考えられる1つもしくはそれ以上の外的基準と，テスト得点を比較することで示される。」
> 「構成概念妥当性は，テストがどのようなものを測っているかを評価することで示される。すなわち，テストにおけるパフォーマンスを，概念や構成概念がどの程度説明できるかである。」

この1966年版の *Standards* では，妥当性はテスト固有の性質ではなくテストの使用目的によって決まること，および，1つの使用場面で1つの妥当性だけが重要になることは稀であることが記されている。しかしそれは複数の種類の妥当性が必要になるという意味であり，*Standards* の本文や個々の規準は，3種類の妥当性ごとに分けて説明されている。

2－4．妥当性の統一理論と S. Messick

構成概念妥当性が妥当性の中心であるという考え方が広く受け入れられようになったのは，1970年代の後半になってからである（Angoff, 1988）。

1970年代といえば，アメリカ合衆国内では，雇用機会の均等に関するガイドラインが出され，採用試験や入学適性試験に関連した訴訟がいくつか起こった時代である（平井, 2005）。その時代，現場では，テスト使用の妥当性をいかに示すかに関心と努力が注がれていたであろうことは，想像に難くない。こうしたときには，1966年版の *Standards* のように妥当性が細かく分かれ，集めるべきエビデンスを示唆してくれる方が便利である。しかし当時は，予測的妥当性，並存的妥当性，内容妥当性，構成概念妥当性，に加えて，収束的妥当性，弁別的妥当性，ドメイン妥当性など，さまざまな名前の“妥当性”が乱立する事態（Newton, 2013, p.305）となり，S. Messick や R. M. Guion らの妥当性の理論家たちは，むしろ構成概念妥当性を基本に据えた妥当性概念の単一化の方向に進んでいった（Kane, 2006）。

> 「構成概念妥当性は，まさに妥当性を統一する概念である。構成概念妥当性によって，基準関連の検討や内容面の検討は共通の枠組みに統合され，その枠組みによって理論から関連性が合理的に仮定され，検証される。･･･（中略），統合の橋渡しとなるのは得点の有意味性もしくは解釈可能性であり，それが構成概念妥当性を検証する目標である。構成概念の意味は予測的関係を仮定するときの理論的根拠となり，テスト内容の適切性と代表性に関する判断の理論的根拠ともなる（Messick, 1980, p.1015）。」

Messick はまた，テスト得点の異なる解釈（得点解釈の前段階には，当然のことながらテストの使用目的が存在する）に必要なのは異なるエビデンスであって，異なる妥当性が必要なのではないと述べ，妥当性は，エビデンスによるテストの解釈と使用についての全体的な正当化の程度だと考えた（Messick, 1980）。また，構成概念妥当性以外のさまざまな“妥当性”は，データ収集やデータ分析の方略やテストの効用，研究エフォート等と見なすべきであり，これらは妥当性の本質部分ではなく，エビデンスの性質を表すものだと主張した（Messick, 1980）。ここで，Messick の考えをまとめるものとして，Messick による妥当性の定義をあげておく。

> “Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on the test scores or other modes of assessment (Messick, 1989, p.13).”

ここで inferences とは，得点から構成概念に対して推論することを指し，得点の解釈とほとんど同じような意味と考えられる。1966年版の *Standards* の記述と比較して，抽象度の高い，包含的な表現になっている点が注目される。妥当性をいくつかのタイプに分けて論じていた時代は，そのタイプに応じてどのようなエビデンスを収集すればよいかが判断しやすかった。し

かしMessickの定義では，妥当性検証において何をどうすればよいか，どこまで追及すべきかなど，その内容や境界は曖昧である。

具体的に考えてみよう。構成概念の理論的把握により，他の変数との関係やテスト内部の構造，テストに用いられるべき素材や回答のさせ方，回答に至る内的プロセス，下位集団差等々に関する予測が立てられる。妥当性検証ではこれらの予測を検証していくのであるが，そのさいのエビデンスには，例えば他の変数との予測的な関係，収束的・弁別的な関係，テスト項目群の因子分析結果，信頼性係数，平均値の下位集団差などの実証的な量的エビデンスのほか，受験者の内観報告やテストの内容分析などの質的なエビデンスまで含まれる。また，テストの作成過程に関する情報やテストを実施した結果も，テスト得点の意味・解釈に関連してくるため，妥当性のエビデンスに含まれる。上記以外にも理論的予測やエビデンスは考えられ，何をどこまで収集すればいいのか，その総合的な評価はどのようにするのか等は，Messickの定義からは判断できず，得点の解釈者が自分で決めなければならない。妥当性概念が統一されたことで，妥当性の扱いは，簡潔になるどころか逆に広がり，曖昧になったということができる。

2－5．*Standards* に表された妥当性の考え方

「妥当性は本質的には構成概念妥当性であり，いくつかの部分に分かれるものではない」という考え方を，“妥当性の単一観 the unitary view of validity”と呼ぶ。この考え方では，妥当性は全体でひとつであるため，「構成概念妥当性」とは呼ばずに，ただの「妥当性」と呼ぶ。

これに対し，それまで主流であった「妥当性は基準関連妥当性，内容妥当性，構成概念妥当性，の3つからなる」とする考え方を，“妥当性の三位一体観 the trinitarian view of validity”と呼ぶ。ここで“三位一体”とは，「三位一体は計量心理学的な救済に至る3本の異なる道を表している」「キリスト教の神学で三位一体とは，神が3人の人間として顕れたことを指す。計量心理学の神学で三位一体とは，妥当性が3つの方向からエビデンスで示されることを指す」（いずれも，Guion, 1980, p.386）と述べたGuionの論文から来ていると思われる。

1974年版の *Standards* では，3つの妥当性がタイプとして表記されており，三位一体観に立っていることがわかる。一方でこの3つの妥当性は操作的にも論理的にも互いに関係しているとも書かれており，単一観的な考え方の兆しがうかがえる（Sireci, 2009）。その次の1985年版になると，「妥当性は単一の概念である（AERA, APA, & NCME, 1985, p.9）」と明記され，妥当性の単一観が前面に出される。1985年版の妥当性の定義は，

> “validity always refers to the degree to which that evidence supports the inferences that are made from the scores (AERA, APA, & NCME, 1985, p.9).”

となっており，前頁で述べたMessickの考え方に沿っていることがわかる。その後，妥当性理論にはあまり大きな変化がなく，1999年版の *Standards* でも，1985年版と同じく，

> “Validity refers to the degree to which evidence and theory support the interpretations of test scores (AERA, APA, & NCME, 1999, p.9).”

と説明されている。エビデンスには主観的な評価や理論的根拠を含めて考えることもあり，その点において，1999年版の考え方は1985年版と大きな違いはないといえる。ただし1999年版では，テスト得点の解釈に言及する箇所ではproposed interpretationや，recommended interpretationなど，基本的に「意図した」という意味の形容詞がつけられ，「テスト得点を解釈するとき，その解釈が真である保証はない。ユーザーはテストの得点はこういう意味であろうと想定して使っているのである。」という考え方がより強調されているようである。テストが生活のさまざまな部分で使われ，テストユーザーの裾野が広がって，同じテストでもいろいろな解釈が行われうるようになった状況も影響しているのであろう。

1999年版には，実質的な変更もある。ひとつは個別の規準の部分で，テストが特定の成果をもたらすと主張する場合の妥当性について明示的に扱うようになった（例えば規準1.22, 1.23, 1.24）点である。

本文部分では，妥当性検証のためのエビデンスに関する記述が根本的に変更された。1985年版は，単一観に立ちながら，エビデンスに関しては読者が混乱しないようにと，三位一体観に則ったカテゴリーに分けて解説していた。一方1999年版はMessickの枠組みを採用し，妥当性のエビデンスを従来とまったく異なる以下の5つの源 sourceに分けて示している。

・内容面のエビデンス
　内容妥当性と呼ばれていたもの，実施手続き，採点手続きなど
・回答プロセス面でのエビデンス
　回答中の思考プロセスの検討や，目の動き，反応時間など
・内的構造面でのエビデンス
　因子分析の結果や信頼性係数など
・他の変数との関係性によるエビデンス
　基準関連妥当性や収束的・弁別的妥当性と呼ばれていたもの，法則論ネットワークなど
・テスト実施の結果によるエビデンス
　構成概念に関係のない部分で特定の集団に得点差が生じることなど

「テスト実施の結果によるエビデンス」は，この1999年版から初めて盛り込まれたものである。テスト実施の結果とは，例えば，測定したい構成概念以外の成分のために一部の回答者が不利になった場合や，テストが構成概念のドメインの一部しかカバーしないために一部の回答者が不利になった場合などである。いずれも，得点の意味・解釈が想定されたものと異なる恐れがあるため，妥当性に疑問が生じることになる。前者は“構成概念に無関係な分散”，後者は“構成概念の代表性不足”と呼ばれ，いずれも測定の非妥当性の大きな要因とされている。

妥当性の結果的側面は，Messick が打ち出した考え方である。テストを使用した結果生じた事態も得点の意味や背景の理論にフィードバックされるべきで，結果的エビデンスも妥当性のエビデンスだと考える。Messick はテスト使用に伴う社会的波及効果まで視野に入れた妥当性を考えていたようであるが（Messick, 1980, 1989, 1995 など），1999年版の *Standards* ではそこまで広げず，本文では測定成分の非妥当性部分に限って盛り込んだようである。この妥当性の結果的側面については，後にまた取り上げる。

2－6．補足

ここまで，妥当性概念の変遷のうち，主流といえる部分を見てきた。その中で触れられなかった点を，2つ補足しておきたい。ひとつは，古典的で広く知られている定義についてである。

> “validity is the degree to which a test measures what is supposed to measure (Garrett, 1937, p.324; Sireci, 2009 から引用).”

同様の定義はその他多くの文献にも見られる。時代的には基準関連モデルと同時期の古いものであり，そのため古典的定義とも呼ばれる。“測ると想定されるもの”を変えれば妥当性が大きく変わってしまう，妥当性をテスト固有の性質としている，などとして，曖昧で不完全だとの批判もあるが，現在でもこの定義を目にすることは多い。1980年代以降の *Standards* や Messick 流の定義が抽象的で複雑なため，一般の心理学研究者にとっては，こちらの方が直観的でわかりやすいのであろう。

二つめは，「構成概念」の意味である。Cronbach と Meehl は，「理論的構成概念」という用語を，その特性を取り巻く理論によって仮定され，意味や役割が定められる，文字通り構築された概念という意味で用いているようである。また「テスト得点を解釈するときに言及する属性 attribute が構成概念である（Cronbach & Meehl, 1955, p.283）」とも述べ，構成概念の例として知能，健忘，快活さ，不安などを挙げている。これらの概念は現在でも「構成概念」として扱われているが，現在の「構成概念」は“現象を説明するために仮定された，直接観察できない心理的な特性”という程度の意味で[5]，潜在変数 latent variable とほとんど同義で用いられることすらある。法則論ネットワークによって概念が規定されるという部分が，現在の「構成概念」の定義にはほとんど含まれていないと考えられるので，注意が必要である。

3．妥当性概念の現在と位置づけ

3－1．最新版における定義

「1．はじめに」でも述べたように，1999年版の *Standards* 以降，その改訂作業と歩調を合わせるかのように，妥当性に関する議論が活発になった。その背景には，テスト使用の現場が拡大し，個人の特性の把握だけでなく，教育プログラムや教育システムの効果測定，行政の判断資料など，テストの活用のされ方が二次，三次と多岐に渡るようになったという状況もあるだろう。テストに期待される役割や使い方が多様化すれば，「妥当性」という概念がどこまでを対象とすべきかについても意見は多様化する。「妥当性」の対象を“心理測定”に限定したい人々がいる一方，副次的効果やプログラムの効果測定まで広げた概念化を求める人々もいる。

そうした多様な背景と声とに対し，2014年版の *Standards* がどのような回答をするのか注目されたが，

蓋をあけてみると妥当性の定義自体には大きな変更はなかった。2014 年版の定義は，

> "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests (AERA, APA, & NCME, 2014, p.11)."

となっており，最後の for proposed uses of tests の部分が追加となっただけである。ある使用目的に伴って得点の解釈が生じるという考え方は 1999 年版でも打ち出されており，今回の変更は従来の定義をより明確にしたものと考えることができる。妥当性の章におけるその他の変更点は，多くが記述の追加や表現の修正であった。具体的には，"ユーザーがある目的でテストを使用することに伴う得点解釈" という点が強調されたり，事例が追加されたり，技術の進化に対応したりしている。

ただし，扱いが大きく変わった部分がひとつだけある。結果的なエビデンスの部分である。結果的側面が初めて取り入れられた 1999 年版では，本文において説明に 1 ページ分のスペースが割かれていた。それが 2014 年版では本文の説明が 2 ページ半に増え，内容においても単なる例の追加などではなく，考え方の範囲が広がった。1999 年版では，構成概念に無関係の測定成分や，構成概念の反映不足に限って結果的側面を取り上げていたが，2014 年版では，得点解釈の延長上に行われる主張（例えば，アチーブメントテストの結果から成績の振るわない学校を見つけるというシステムについて "テスト結果は，学習向上につながる" と主張すること）や，テスト実施により意図せざる影響が出ること（例えば，小論文テストをコンピューターで自動採点した結果，測定の一貫性は高まったが教師の採点スキルが低下したこと）が結果的側面の例に挙げられている。こうした変更は 1999 年版の規準部分で簡単に扱った二次的な機能の部分を他のエビデンスと同等に扱うものであり，実質的には「妥当性」概念の適用範囲を結果の方向に広げたものだといえる。つまり 2014 年版は，妥当性の定義自体は大きく変えなかったが，よりテストの現実的側面を取りこむようになったのである。

3－2．整理と位置づけ

ここでこれまでの妥当性理論の歴史を整理し，妥当性の概念がカバーする範囲という視点からとらえ直してみたい[6]。

①構成概念の特定（測定したい特性の定義）
②測定手続き（素材や用紙・用具，実施方法，得点化ルールなど）
③テストの使用目的（どういう機能を期待して用いるか）
④得点解釈（解釈者が想定する得点の意味）
⑤解釈に基づく使用（テスト結果に基づく何らかの判断，行動）
⑥テストの効用（機能や有効性の評価）
⑦使用の結果生じること（想定範囲内，想定範囲外）

このうち，古典的定義では①に対する②の適切さを取り上げているが，テストが求める機能を果たすかどうかという視点も盛り込まれているため，③まで含んでいると考えられる。Cronbach と Meehl による構成概念妥当性の考え方では，法則論ネットワークによって構成概念が規定されるため，①の適切さと②の適切さが同時かつ相互作用的に評価される。ただし Cronbach と Meehl は，構成概念妥当性の検証を "テスト得点をこう解釈したいが，それをどのように弁護すればよいか（Cronbach & Meehl, 1955, p.284）" であるとしていることから，彼らの構成概念妥当性は解釈まで含むもの，すなわち①から④までと考えることができる。妥当性の三位一体観は，3 つのタイプの妥当性が住み分けているものの，基本的には得点解釈の妥当性を考えているため，これも①から④までを対象としているといえる。妥当性の単一観に立つ 1985 年版の *Standards* は，解釈が対象となることが明示されており，同様に①から④までを対象とした概念化といえる。1999 年版の *Standards* の定義は，①から④に重点があるものの，⑥や⑦も少し取り入れられている。Messick は①から⑦までを考え，中でも⑥や⑦に関する価値判断を重視していたように読める。2014 年版の *Standards* では，Messick ほど⑥と⑦を重視してはいないものの，本質的には①から⑦までを広く想定しているといえる。

4．日本の心理学研究への適用

ここまでの議論からわかるように，現在主流となっている妥当性概念には，アメリカ合衆国のテスト事情が色濃く反映されているとみることができる。そこでは，国や地方の教育行政や産業でのテストおよびテス

ト結果の利用や社会的波及効果などをにらみながら，テストの使用が適切であることをエビデンスによっていかに論証するかに重点が置かれているように見える。そうした実用場面重視の概念化を，我が国の心理学研究にどのように適用できるのだろうか。

日本の心理学研究で測定の妥当性が関わる場面は，多くが心理尺度の作成・使用の場面であろう。心理尺度は多用されているといってよい（内田, 2012）。心理尺度を用いた研究には，他者の作成した尺度をそのまま用いるもの，新たな心理特性を定義して尺度を作成するもの，既存の尺度の改良版を作るもの，外国の尺度の日本語版を作成するもの，既存の尺度から項目を借用・改変・組合せて新たな尺度とするもの，より少ない項目数で尺度を構成するものなど，これらのバリエーションや組合せを含めて多種多様である。中には他の研究者によってその後何度も使われる尺度もあるが，その研究限りで使い捨てられる尺度もある。このような尺度の使用状況において，「妥当性」の概念はどのように考えればいいのだろうか。

ただし本論文は，新たに日本の心理学研究用の定義を提案しようというのではない。*Standards* が実質的に世界標準であることを鑑みると，別な定義を提案しても心理学研究者にとって混乱を招くだけである。妥当性の定義はそのままに，研究において留意すべき重要な部分とそうでない部分の濃淡を検討することを目的とする。よって以下の議論では，2014 年版の *Standards* にならい，妥当性を「提案されるテストの使い方において，得点の解釈がエビデンスと理論によってどの程度支持されるか」を意味するものとする。以下，大きく 2 つの場合に分けて検討する。

尺度が基本的に新作とみなされる場合（改変，翻訳，組合せ，短縮版などを含む）

作成された尺度の使われ方としては，集団として統計的に分析する場合と，個人の特徴を記述するために用いる場合とがある。前者は，多人数に実施し，他の変数と絡めて集団として分析する研究である。尺度得点にもとづいて群分けし分散分析にかける場合も含まれる。後者は，ケース研究などで個人の特徴を記述するために，背景情報とともに知能検査得点や人格検査などの得点を付記する場合などである。こうした研究場面における尺度の作成作業から使用までを想定して，前節に挙げた，妥当性の概念がカバーする範囲に照らして検討する。

まず“①構成概念の特定”であるが，これは尺度を作成・使用する前に行うべき必須の作業といえる。この意識が希薄な研究としては，例えば探索的因子分析によって尺度を構成したときなど，尺度項目が先に決まり測られる特性が後から“命名”される場合があげられる。測られる特性が先に決まらないと，尺度得点が想定する構成概念の測度として適切かどうかの議論ができない。測定したい構成概念の特定（定義）は，何のためにその尺度を作成・使用するのかという問題に直結する。つまり，研究の目的やリサーチクエスチョンに組み込まれた，研究の意義に関わる問題といえる。

“②測定手続き”も，重要な部分といえる。同じ質問文でも，評定尺度に付記される数値ラベルが異なったり，直前の項目が異なったりすると，その項目の回答分布が異なることがある（Schwarz, 1999）。対象者集団が異なればなおさらである。自分の測定したい特性を最も忠実にくみ取ることのできる測定手続きは，先行研究と自分の研究では異なるかも知れないのである。自分の研究においてどのような測定手続きが望ましいかは，必ず検討しておきたい。

“③目的”は，心理学研究ではたいていの場合「測りたい特性の測度とするため」となり，他の使用目的はほとんど考えられない。先にも述べたように，この部分はリサーチクエスチョンに関わる重要な部分といえる。

“④得点解釈”は，基本的に「特性の定義どおり」となるが，得点が実際に定義通りの内容を反映しているかどうかの検討は，それとは別に必須である。測りたい特性と無関係な成分が得点に大きく混入していないかどうか，測りたい特性の定義ドメインが過不足なくカバーされているか，測定誤差は大きくないか，などを検討しておきたい。

“⑤使用”と“⑥効用”は，集団として分析する研究ではほとんど関係ないといってよい。最後の“⑦結果”は，各種の倫理規定に従う限り，テスト実施による回答者へのネガティブな影響は極力抑えられる。ポジティブな影響があったとしても，得点解釈や概念定義にフィードバックさせる必要があるほどの影響は考えにくい。ただし，尺度得点にもとづいて個人の処遇が決定され，その後異なる介入を行うような研究では，“⑤使用”，“⑥効用”，“⑦結果”の評価も行う必要があるだろう。

このように考えると，研究で新たな尺度を作成・使用する場合は，“①構成概念の特定”と“③目的”の議論をきちんと行い，それに適切に対応できる“②測定

手続き”になっているかどうかを検討し，“④得点解釈”がエビデンスによって十分支持されているかどうかを検討することが，大きな部分を占めるといえる。

他者の作成した尺度をそのまま用いる場合

市販の尺度であったり，論文で発表されていたりするものを，項目セット，項目の表現，項目の順番，回答のさせ方等，何も変更せずにそのまま用いる場合を想定する。集団として統計的に分析する場合と，個人の特徴を記述するために用いる場合の両方がありうる。

これらの場合，尺度が作成された時点で，“①構成概念の特定”と“②測定手続き”の検討がかなりの程度済んでいる。その尺度が開発されたときの回答者集団と同質の対象者に実施し，同じ測定手続きと使用目的で，同じ解釈のし方をするのであれば，尺度の使用者が行うことはあまり多くない。基本的には，尺度の作成者側がこれらの検討を十分に行ったかどうかの確認と，その尺度を使用した他の研究を概観し，自分の使用目的と対象者集団に対して尺度が有効に機能するかどうかの確認をすれば，基本的に間に合うのではないだろうか。一方，自分の使用場面が，尺度が開発されたときの状況と何らかの部分で異なる場合は，その部分に応じて，“①構成概念の特定”，“②測定手続き”，“③目的”，“④得点解釈”の該当部分を使用者側が検討する必要がある。“⑤使用”，“⑥効用”，“⑦結果”に関しては，尺度が基本的に新作と見なされる場合と同様に考えればよい。

このように考えると，尺度が新作であろうと既存のものであろうと，研究における尺度の作成・使用者は，“①構成概念の特定”から“④得点解釈”までに重点を置いて考えればよいことになる。Cizek (2012) は，得点の解釈とテストの使用を分けて考え，妥当性の概念を得点解釈までに限定し，その後のテスト使用は“適切性”で議論すべきだと主張した。Cizek のこの主張は，心理学研究で妥当性を考えるときのひとつの参考になるといえるだろう。

妥当性理論がわかりにくい理由のひとつに，妥当性と妥当性検証とが絡み合って議論される点がある。本論文では主に妥当性の概念定義に焦点を当て，その歴史的経緯を踏まえて，妥当性の定義とそのカバーする範囲という考え方を導入した。しかし妥当性がカバーする範囲は，すなわちエビデンスを示すべき範囲でもある。本来は同時に妥当性検証についても論じる事が望ましいが，エビデンスは多岐にわたり，妥当性の論証自体も丁寧に扱う必要がある。具体的な妥当性検証の手続きについては次の論考のテーマとしたい。

文献

赤木愛和・池田央（監訳）(1993). 教育・心理検査法のスタンダード　図書文化 (AERA, APA, NCME, 1985, *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.)

American Educational Research Association & National Council on Measurements Used in Education. (1955). *Technical recommendations for achievement tests*. Washington, DC: National Education Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological*

tests. Washington, DC: American Psychological Association.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.) *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates. Pp.19-45.

Bennett, R. E. (2012). Consequences that cannot be avoided: A response to Paul Newton. *Measurement: Interdisciplinary Research and Perspectives*, 10, 30-32.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. UK: Cambridge University Press.

Borsboom, D., Mellenbergh, G. J., & Heerden, J. van (2004). The concept of validity. *Psychological Review*, 111, 1061-1071.

Braun, H. I., Jackson, D. N., & Wiley, D. E. (2002). *The role of constructs in psychological and educational measurement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Chatterji, M. (2013). *Validity and test use: An international dialogue on educational assessment, accountability and equity*. UK: Emerald Group Publishing.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, 17, 31-43.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Fiske, D. W. (2002). Validity for what? In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.) *The role of constructs in psychological and educational measurement*. Mahwah, NJ: Lawrence Erlbaum Associates. Pp.169-178.

Garrett (1937). *Statistics in psychology and education*. New York, NY: Longman Green.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6, 427-439.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, 11, 385-398.

平井洋子 (2005). 公平性に対する測定論的アプローチの歴史的展望－選抜テストにおけるマイノリティの低得点を背景に－ 東京都立大学人文学報 第358号 1-29.

平井洋子 (2006). 測定の妥当性からみた尺度構成－得点の解釈を保証できますか 吉田寿夫（編）心理学研究法の新しいかたち 誠信書房 Pp.21-49.

Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology*, 10, 93-98.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement*, 4th ed. Westport, CT: Praeger. Pp.17-64.

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age. Pp.39-64.

Lane, S. (2012). Consequences of assessment and accountability systems are integral to the argument-based approach to validity. *Measurement: Interdisciplinary Research and Perspectives*, 10, 71-74.

Lissitz, R. W. (2009). *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age.

Loevinger, J. (1967). Objective tests as instruments of psychological theory. In D. N. Jackson & S. Messick (Eds.) *Problems in human assessment*. New York, NY: McGraw-Hill. Pp.78-123. (reprinted from *Psychological Reports*, 1957, Monograph Supplement 9).

Markus, K. A. & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*. 35, 1012-1027.

Messick, S. (1989). Validity. in R. L. Linn (Ed.), *Educational measurement*, 3rd ed. New York: American Council on Education / Macmillan. Pp.13-103.（メシックS. 妥当性 池田央・柳井晴夫・藤田恵璽・繁桝算男（監訳） 1992 教育測定学原著第3版 上巻 みくに出版 Pp.19-145.）

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*. 50, 741-749.

村上隆 (2003). 測定の妥当性 日本教育心理学会（編） 教育心理学ハンドブック 有斐閣 Pp.159-169.

村山航 (2012). 妥当性－概念の歴史的変遷と心理測定学的観点からの考察 教育心理学年報 第51巻 118-130.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*, 10, 1-29.

Newton, P. E. & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, 18, 301-319.

Newton, P. E. & Shaw, S. D. (2014). *Validity in educational & psychological assessment*. UK: Cambridge Assessment.

Plake, B. S. & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME *Standards for Educational and Psychological Testing? Educational Measurement: Issues and Practices*, 33(4), 4-12.

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review*, 16, 290-296.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54(2), 93-105.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.) *The Concept of validity: revisions, new directions, and applications*. Charlotte, NC: Information Age. Pp.19-37.

内田照久（2012）．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．教育心理学年報 第51集 63-72.

Wainer, H. & Braun. H. I. (1988). *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D. & Chan, E. K. H. (2014). *Validity and validation in social, behavioral, and health sciences*. Switzerland: Springer International Publishing.

注

1 以下，本論文では，検査や心理尺度，試験，アセスメント等を含む広い概念として「テスト」という用語を用いる。

2 1999年版の改訂作業は，テスト会社や企業，NPOからもメンバーを入れ，以下のスケジュールで行われた。2005年，改訂の必要性を検討する委員会を設置。2008年，改訂作業を行う委員会を設置し，パブリックコメントを募集。2011年，改訂版の草稿を公開しコメントを募集。2012年，再改訂した原稿を法的チェックにかける。その後3学会にて承認を受け，2014年出版（Plake & Wise, 2014）。

3 例えば，ある学校の入学試験の得点でも，一般的な学力，その時点までの達成度，その学校における学業適性など，さまざまな解釈がありうる。そのため，解釈ごとに妥当性は異なると考える。

4 1966年版では，予測的妥当性と並存的妥当性が基準関連妥当性にまとめられている。

5 例えば1999年版と2014年版の *Standards* では，巻末の用語集で construct を "the concept or the characteristic that a test is designed to measure" としか説明していない。

6 この分類は暫定的なものであり，数字も便宜的に割り当てたものである。