論　説

構成概念妥当性の検証方法に関する検討
――弁別的証拠と法則的証拠を中心に――

中　村　陽　人

1．はじめに

妥当性とは，尺度が測定しようとしているものを実際に測っているかどうか，その程度を表す概念であり（吉田 1994），ゆえに，尺度にとって妥当性という裏づけは欠くことのできない必須の要件である．そのため，尺度開発に関わる研究はもちろんのこと，構成概念を扱うような定量的研究において妥当性の検証は非常に重要なものである．

しかし，これまで数多くの妥当性が提案され研究されてきたが，研究によって用いられる妥当性の種類が異なったり，同じ妥当性でもその検証方法が様々であったりすることが多い．もちろん妥当性を示すことが目的であるから，その方法が様々であることに問題はないが，その選ばれ方が単に計算しやすさなど恣意的であったり，あるいは過去の研究に従って盲目的に選択されたりしたものが多く，妥当性が本来の役割を十分に果たしていないことが多いと考えられる．

そこで本稿では，まず妥当性概念がいかに変遷してきたのかを辿り，その具体的な検証方法についてまとめる．次に母相関係数の検定や推定を用いた弁別的証拠の検証力について吟味する．また，尺度開発など一つの構成概念に対して多くの項目を用いた研究における法則的証拠の検証について吟味する．

2．妥当性概念の変遷

次章で見るように，研究分野によって妥当性の枠組みは多少異なる．しかし，妥当性に関して研究されてきた時間の長さや，妥当性への関心の高さは圧倒的に心理学分野が抜きに出ている．村上（2006）が「妥当性は心理測定上，最も重要な概念であるので，歴史的にもかなり古くから議論されてきた」と述べているように，心理学分野では伝統的に心理尺度を作成することが多く，そのために妥当性の議論が必要不可欠であったことがその原因と考えられる．よって心理学分野における妥当性概念の枠組みの変遷を中心に据えて検討することにする．

心理学分野においては，米国教育学会（AERA, American Educational Research Association），米国心理学会（APA, American Psychological Association），全米教育測定協議会（NCME, National Council on Measurement in Education）が合同で，専門家に対する勧告や基準として

妥当性の概念を定義している．APA（1954）の専門的勧告では，内容妥当性，予測妥当性，併存妥当性，構成概念妥当性の計４つの妥当性が示されている．これらはAPA（1966）になると，予測妥当性と併存妥当性が統合され，新たに基準関連妥当性となっている．つまり，内容妥当性と基準関連妥当性と構成概念妥当性の３つの妥当性が並立的に位置づけられていたのである．

内容妥当性は，主に複数の専門家による項目内容の検討が行われる[1]．ただし内容の検討だけでは妥当性の証明にはならず，実際のデータの収集とその分析が必要になる（村上 2003）．具体的に言えば，尺度内容を検討するために測りたいドメインと尺度得点との対応を考えることになる．つまり内容妥当性をつきつめると構成概念妥当性に行き着くことになる（平井 2006）．また，基準関連妥当性も，何が外的基準として適切かを判断するためには，尺度が測定しようとする構成概念はどのようなものかを考える必要がある．さらに，外的基準との関係が正しくデータに反映されるためには構成概念と項目との対応もしっかりとできていなければならない．結局，基準関連妥当性も構成概念妥当性に行き着くことになる（平井 2006）．こうして，1980年代になると徐々に構成概念妥当性の拡大解釈が進み，構成概念妥当性は基準関連妥当性や内容妥当性を含む上位概念であると考えられるようになった．1985年の基準では伝統的な妥当化の手段を，内容に関連した妥当性の証拠，基準に関連した妥当性の証拠，そして構成概念に関連した妥当性の証拠として位置づけた．さらに構成概念妥当性の拡張の議論は進み，Messick（1989）ではテスト使用の倫理的ないし社会的結果までを含めるべきであるとし，Messick（1995）では，内容的側面，本質的側面，構造的側面，一般化可能性の側面，外的側面，結果的側面の6つの証拠が，妥当性を示す要件としてあげられている．この倫理的・社会的結果を妥当性概念に含めるかどうかは異論のあるところで，Popham（1997）は倫理的・社会的結果の重要性を認めた上で，それを妥当性の概念に含めることについては，概念を拡散させ混乱を生じさせると批判的である．また，村上（2006）も評価方法がほとんどないことなどから，社会的重要性と妥当性の問題は切り離したほうがよいとしている．しかし，AERA, APA, & NCME（1999）でも，５つの妥当性の証拠の根拠[2]（sources of validity evidence）をあげ，結果に関する要素を含んでいる．これまでのところ妥当性概念に関する共通見解は得られず，研究者の間でも妥当性概念の混乱は続いているが，それはあくまで分類や各要素の位置づけの違いであって，示すべきものに大きな差はない．妥当性をどのように捉えるにしても，その検証には唯一の方法というものがあるのではなく，複数の観点から検証しなければならない，というのが多くの研究者に共通した見解といえる（清水 2005）．

鈴木（2005）は妥当性の概念分類の歴史的推移を整理し，表１に示すようにまとめている．

消費者行動の分野においても研究で心理変数が扱われることは多く，厳密さの程度の差はあれ，構成概念妥当性の検証が行われてきている．阿部（1987）は構成概念妥当性を示すための必要条件として，信頼性，収束妥当性，弁別妥当性，法則的妥当性の4条件をあげている．法則的妥当性は構成概念間の法則的な関係性を扱うもので基準関連妥当性を含むと考えられるが，それが構成概念妥当性の必要条件とされている点は現在の妥当性の考え方（構成概念妥当性に統合され，各妥当性はその証拠であるとする考え方）に通じる．前章の内容的側面に関わるものがないこと以外，ほぼ重ねることができる．したがって，心理学分野での現在の枠組でも阿

---

[1] 専門化の判断を数値化して，その間の一致率や相関を調べるといった方法がとられることもある．

[2] テスト内容（test content），反応プロセス（Response processes），内部構造（internal structure），他の変数との関係（relations to other variables），テスト得点化の結果（consequences of testing）である．

表１　妥当性の概念分類の歴史的推移

年代	1950		1960-70		1980		1990	
出典	APA（1954）		APA, AERA, NCME（1966; 1974）		AERA, APA, NCME（1985）		AERA, APA, NCME（1999）	
分類	1	content validity	1	content validity	1	content-related evidence of validity	1	evidence based on test content
							2	evidence based on response process
	2	construct validity	2	construct validity	2	construct-related evidence of validity	3	evidence based on relations to other variables
							4	evidence based on relations to other variables 4-1 convergent and discriminant evidence
	3	predictive validity	3	criterion-related validity	3	criterion-related evidence of validity		4-2 test-criterion relationships 4-3 validity generalization
	4	concurrent validity						
							5	evidence based on consequences of testing

［出所：鈴木（2005, p.62）］

部（1987）で示された枠組でも，実際の手順にそれほど差があるわけではないことがわかる．ただし，消費者行動の分野では心理学分野に比べて妥当性の検証が疎かにされているものが多い．特に弁別妥当性についてきちんと検証したものは少なく，法則的妥当性にまで踏み込んだものは極々少数である．

なお，構成概念妥当性は他の妥当性の上位の概念として位置づけられるようになったため，妥当性という語と構成概念妥当性という語の差異が区別されずに使われるようになっている．しかし，これまで議論の対象としてきた「妥当性」という語は，本稿の始めに示した吉田（1994）の定義のように「測定の妥当性」という狭義の意味を示すものである．そもそも妥当性という語は「構成概念そのものが妥当かどうか」という意味をも含む広義の語である．したがってこれ以降，広義の意味の妥当性を「妥当性」，測定の妥当性を示す狭義の意味の妥当性を「構成概念妥当性」と表すことにする．もちろん，本稿で対象としているのは構成概念妥当性のことである．

３．構成概念妥当性検証の枠組み

Messickは前章で示したAPAなどによる「スタンダード」の作成で中心的な役割を果たし，その考え方は「スタンダード」の中でも大いに反映されている．また以下で見るように，部分的に議論を呼んでいるところもあるが，総じて緻密な概念の検討と分類がなされていると考えられる．さらに，Messickの分類は構成概念妥当性を広く捉えているため網羅的であり，重要な漏れを防ぐことができると考えられる．したがって，Messick（1989；1995）の6つの要件を中心に，特に尺度開発の場合を想定して構成概念妥当性検証の枠組みをまとめる．なぜなら，最も構成概念妥当性の厳密さが要求されるのは尺度開発の場であるし，構成概念を扱った研究

において，測定項目を選択し，決定するというプロセスは，尺度開発に準じたものであることが求められているからである．

3.1　内容的側面（content aspect）からの証拠

測定したい構成概念に含まれる要素，含まれない要素を明確に線引きしたドメインを定義し，尺度内容がドメインに対応しているか，十分に代表しているかを示す証拠である．つまり，ドメインに関する仕様書[3]，項目の適切性，項目の代表性などを吟味することになる（平井 2006）．尺度の開発者がこれらの吟味を行う場合と複数の専門家に判断を委ねる場合があり，さらに数値化して判断するものと議論による合意で判断するものとがある．数としては，数値化したものは少なく，専門家，あるいは開発者の主観的な判断でなされることが多い（村上 2006）．

項目抽出段階では，尺度開発のベースとなる数多くの項目を列挙しなければならない．尺度開発者の主観的な方法としては，以前作成された項目プールを利用する，質的調査などの結果から考えられる項目を列挙する，仕様書に基づいて考えられる項目を列挙するなどの方法がある．複数の専門家の判断に委ねる方法としては議論による合意で列挙する方法がある．また，機械的に項目を作成する方法があり，Messick（1989）では，写像文（mapping sentence）または項目生成規則（item-generation rule）を利用する方法（Hively, Patterson, & Page 1958）や，一定の形式構造を持つように特定化された1つの項目形式を，置換リストに沿って要素を置き換える方法（Osburn 1968）などが紹介されている．しかし，Hivelyらの項目生成規則は，数学の計算テストのように概念枠組みや問うべき要素が明確なものに対して，どの項目を選ぶかということを対象としている．また，Osburnの方法も概念枠組みは明確なものが対象で，固定された項目形式にどのような表現をあてはめるかが関心の向かう先である．しかし，消費者行動分野の多くの研究の対象とするような構成概念はその枠組みが非常に不明瞭であり，確かに尺度開発においてドメインをできる限り明確に定義することを出発点とはするものの，ある程度の曖昧さは内包しているため，ドメインからもれなく必要な要素を取り出せるかどうかに関心が向けられている．したがって，ここで挙げた機械的な項目作成方法は，消費者行動分野ではあまり有効ではない．

項目選択段階では，データを基にした定量的な方法が用いられることが多いものの，開発者が仕様書などに沿って内容的に判断することもある．また，複数の専門家にテスト項目と内容との関連性を数値評価してもらい，専門家間の一致率や相関係数を調べる方法もある．

上記のように内容的側面からの証拠として示せるものは多くあるが，いずれも程度の差こそあれ主観的な方法であり，証拠としての説得力は高くない．つまり，内容に関連した証拠はそれだけで成り立つのではないことは明らかである．よって，Messick（1989）が述べているように，統一的妥当性の枠組みの中で構成概念と関連付けられた証拠と，うまく協調して機能するかについて確かめる意味で検討するのがよいと考えられる．

3.2　本質的側面（substantive aspect）からの証拠

項目選択の際，内容面からのアプローチにおいては，項目はドメインの領域の特定だけを基礎としてテストに含まれるか否かが決められる．一方，厳密な経験的アプローチにおいては，

---

[3]　尺度開発の始めに構成概念のドメインを明確に規定し，それを文書化したものである．

項目の等質性や因子負荷量といったテスト内部データや，基準との項目相関，基準グループの間の弁別性といった外部データを基礎として項目がテストに含まれるか否かが決められる．この両者の対立の中で構成概念妥当性が持つ重要面を統合しようとするのが本質的側面になる．簡単に言い換えるならば，データを基にした項目選択と内容を基にした項目選択の結果の相違点を解釈し，理由づけを行うということである．

実質的アプローチは２段階に分けることができる．第１段階は，項目プールの拡大と項目分析である．ここでは意図的に対象とする構成概念以外の項目まで広く含めた項目プール内の，各項目の項目分析を行う．項目ごとの反応の分布，平均・標準偏差，項目間相関，G-P[4]分析などの結果から検討することができる．さらに反応プロセスを説明するために，回答時間や反応時間のパターン，プロトコル分析[5]，眼球運動の分析なども行われる．ただし，この場合，これらの手法は主に応答時間や誤謬率から影響を受けた項目困難度（項目の難易度）が基になっている点に注意が必要である．項目困難度はもともと教育心理学におけるテスト理論を基にしているため意味を持つが，心理尺度を対象とする場合，項目困難度というものは項目の回答しやすさのようなものであり，項目の本質からは離れることになる．

第２段階では，第１段階での結果などを基にして，すべての項目に理論的な説明を加える．その際，採用された項目だけでなく，取り除かれた項目に関しても説明を加えるべきである．

また，実質的アプローチを効果的に仕上げるために，複数の構成概念の測定を同時に行うことが有効である．回答者にいくつかの構成概念を表す項目が混在する項目プールを用いることで，収束的証拠や弁別的証拠を同時に提出することができるからである．

3.3　構造的側面（structural aspect）からの証拠

得点の内的構造が構成概念の下位領域や次元性などの理論的構造に一致していることを示す証拠である．得点化手続きの適切性や項目間の相関関係の他，因子分析の結果や一次元性の確認，内的整合性のデータも証拠に含まれる（平井 2006）．

まず，得られたデータから探索的因子分析を行い，下位尺度化する．この結果が理論的構造に一致していれば，構造的側面からの強い証拠となる．しかし多くの場合，探索的因子分析の結果は理論構造と一致しない．このような場合，理論構造に基づいた確認的因子分析を行い，その結果を重視すべきである．なぜなら，探索的因子分析が得られたデータにおける変数間の相関関係のみから，データ主導的に共通する因子を抽出するものであるのに対して，確認的因子分析では，あらかじめ構成概念について観察変数との間の測定モデルを特定化し，そうした因子構造が成立しているのか否かを，与えられたデータでチェックするという考え方に立っているためである（阿部 2002）．次に，次元ごと探索的因子分析と内的整合性による α 係数を算出し，一次元性の確認を行う．ここで，重相関係数の平方[6]（SMC：squared multiple correlation）や

---

[4]　総スコアに従って回答者を３ないし４分割し，最上位（最も好意的）グループと最下位（最も非好意的）グループとで各項目の平均値などを比較するもの．平均値が大きく異なる項目が，弁別力のある良い項目であると判断される．

[5]　被験者に何らかの行動を行ってもらい，その過程で考えていることをすべて口に出してもらう方法．詳細は，阿部（1981）や阿部（1984）を参照のこと．

[6]　SMC：因子分析における共通性の推定値の１つとして用いられるもの．変数j以外のn－1個の変数から推定された変数jの共通因子成分と変数jとの相関係数は，変数jとn－1個の変数との重相関係数に他ならない．したがってある変数の共通性の推定値は，当該変数と他の変数との重相関係数の平方によって与えられる．

修正済み項目合計相関[7]（Corrected Item-to-Total Correlation）の値を利用して項目を吟味する．

また，得点化について考えるとき，各次元レベルの得点と，さらにそれらをまとめた合成得点（composite score）を扱うことが多い．各次元レベルであれば，そこに含まれる項目の得点を単純に足したり，あるいは平均をとったりすることで得点化される．しかし合成得点は各次元の「寄与」を求める必要がある．この寄与を求める方法として，２つの方法が挙げられている．１つ目は，テスト全体がカバーする内容領域が各次元にどのように配分されているかを基にする方法である．この方法は客観的に配分を決定するのが困難である．２つ目は，データから重みを算出する方法である．算出方法はいくつか提案されており，尺度によって算出方法は異なる．

3.4　一般化可能性（generalizability aspect）の側面からの証拠

得点の意味や測定論的特性（平均や標準偏差，項目間の相関構造など）が，ある特定のデータセットだけでなく他の被験者集団，実施場面，実施時期，項目セットに対しても不変であるという証拠である（平井 2006）．すなわち別なサンプルをとるか，大きなサンプルを二分割するなどして，因子や因子負荷の安定性を確認する交差妥当化（cross-validation）の手続きである．

さらに一般化可能性理論による分散成分の検討結果が証拠となるが，再検査信頼性や代替検査信頼性，α 係数も証拠となる．これまで一般化可能性理論はその計算の複雑さが問題となっていた．現在ではSASのVARCOMPプロシジャなど専門のソフトがある他，AmosなどのSEMの実行可能ソフトがあれば一般化可能性係数や多変量一般化可能性係数を推定するのに必要な分散成分や共分散成分が，確認的因子分析モデルによって推定可能であることが示されている（中村 2003）．

3.5　外的側面（external aspect）からの証拠

他の変数との間に理論上想定される相関パターンが実際にも示されるという証拠である．同じ特性を測定していれば相関は高くなるはずであり（収束的証拠），異なる特性を測定していれば相関は低くなるはずである（弁別的証拠）．また，同じ特性を測っていても，測定方法が異なればそれだけ相関は低めになるはずである．多特性・多方法行列やSEMモデルは，こうした予測を体系的に検証するための手法として非常に有効である（平井 2006）．

収束的証拠を示す要件としては，「各構成概念とその指標間に有意なパス係数が存在しその符号が全て正であるかどうか（趙，兪 2004）」「ある構成概念とそれを説明する指標との因子負荷量が統計的に有意かどうか（阿部 1987）」「確認的因子分析において，構成概念とそれを構成する項目間の相関が十分に大きいかどうか（畑井 2004；久保田 2006）．具体的には因子負荷量がその標準誤差の２倍以上であるか，あるいは因子負荷量が0.5以上であるか」などが用いられている．

弁別的証拠を示す要件としては，母相関係数の検定や推定を用いる方法として，「ある構成概念を説明する複数の構成概念間のパス係数が1.0と有意差をもつかどうか」「構成概念間の相関を表す係数の99％信頼区間（各相関係数±2.58SE）が完全な相関である1.0（あるいは－1.0）を含むかどうかの可否を，確認的因子分析を用いて検討し，完全な相関を含むものがなければよい」などの基準がよく用いられている．しかし，この方法は次章に見るように検証力に問題があると考えられる．その他の方法としては，「それぞれの概念によって説明される分散部分が，概念

---

[7]　修正済み項目合計相関：特定の項目の得点と，その項目以外の項目の合計得点との相関係数．

間の相関の２乗よりも大でなければならない（阿部 2002)」「仮説モデルと，それより少ない因子を持つモデルとを比較し，そのχ^2値の差を検定することでモデルの妥当性を検証する[8]（川上 2005)」などがある．しかし，後者は因子の数が増えると組み合わせが格段に増える（例えば，６因子になるとその組み合わせは100を超える）ため，この方法は因子数が少ないときに有効な方法と言える．

法則的証拠を示す要件としては，「各構成概念間の因果関係を示すパス係数が有意かどうか（趙，兪 2004)」が用いられている．ただし，法則的証拠を示した研究の数は非常に少ない．また第５章で見るように，尺度開発など，一つの構成概念に対して数多くの項目を尺度として用いる場合，全体のモデルを基に法則的証拠を吟味することは行われていない．それはモデルが非常に複雑になるためにモデル全体の適合度指標の値がかなり低いものとなることが大いに関係していると考えられる．

3.6　結果的側面（consequential aspect）からの証拠

妥当性は，特定の状況でその尺度を使用することの適切さを問題にする．したがって，その尺度を使用した結果生じた事態についても，妥当性評価の対象に組み込まれる．例えばある尺度を用いた結果，ある下位集団が系統的に不利になったとすれば，その場面でその尺度を使用したことが適切だったとは言い難い．短期的・長期的な悪影響が理論的・経験的に生じない，もしくは予見されないという証拠が必要である（平井 2006)．

しかし，Messickのこの考え方はあまりにも複合的で多岐にわたる問題を内包しているため，実際にはあまり採用，実践されなかった（鈴木 2005)．また，Popham（1997）はテストの社会的重要性は測定値の妥当性に無用な混乱をもたらすだろうと主張した．さらに村上（2006）も社会的重要性と測定値の妥当性の問題は切り離した方がよいとしている．少なくとも現段階においては，結果的側面は除いて考えるのがよいと考えられる．

これまでの議論をまとめると，測定の妥当性を表す上位概念として構成概念妥当性を置き，その５つの側面を表す証拠として，Messick（1995）の６つの要件から結果的側面を除いた，内容的側面，本質的側面，構造的側面，一般化可能性の側面，外的側面を考えるのがよいと考えられる．それぞれの側面について具体的な手順をまとめたものが表２である．

４．弁別的証拠に関する検証方法の検討

前章でもふれたように，弁別的証拠を示す際に母相関係数の検定や推定を用いる方法がよく使用されている．しかしながらその検証力に関して疑問が残る．以下，実際にいくつかの値を用いて検討してみることにする．

---

[8]　仮説モデルとのχ^2値の差がすべて有意であればよい．仮説モデルがｎ因子モデルなら，ｎ−1，ｎ−2，・・・，１因子モデルのすべての組み合わせを考えることになる．例えば仮説モデルが（A，B，C）という３因子モデルの場合，２因子モデルとして（A+B，C)，(B+C，A)，(C+A，B）の３つが，１因子モデルとして（A+B+C）の計４つのモデルが得られ，それぞれのモデルのχ^2値と自由度から検定を行うことができる．

表２　構成概念妥当性の検討手順

構成概念妥当性				
内容的側面	本質的側面	構造的側面	一般化可能性の側面	外的側面
○ドメインを明確に定義し，仕様書にする ○項目の抽出法に関する吟味［以前作成された項目プールの利用，質的調査の結果，仕様書に基づく列挙］ ○専門家による項目の適切性と代表性の吟味［一致度などで数値化する，議論による吟味］ ○尺度開発法あるいは尺度選択の正当性の検討 ○サンプルの正当性の検討 ○分析の正当性の検討	○項目分析［項目ごとのデータの分布，平均，標準偏差，項目間相関，G-P分析］ ○項目プールの拡大［複数の構成概念の測定項目を同時に問う］ ○反応プロセスの確認［回答時間，反応時間，プロトコル分析，眼球運動など］ ○項目の採否に関する内容の検討［全項目について採否にかかわらず検討を加え，残した理由と落とした理由を明確にする］ ○収束的証拠［外的側面参照］ ○弁別的証拠［外的側面参照］	○全体の探索的因子分析の結果と理論構造の比較 ○一次元性の確認［構成概念あるいは次元ごとの探索的因子分析，α 係数の算出，重相関係数の平方，修正済み項目合計相関の算出］ ○得点化の仕方［合成得点の算出方法，重みづけの方法など］	○交差妥当化［サンプルを二分して一方でモデル作りや予測を行い，もう一方で確認を行う］ ○一般化可能性理論による分散成分の検討	○収束的証拠［ある構成概念とそれを説明する指標との因子負荷量が統計的に有意かどうか，標準誤差の２倍以上であるかどうか，など］ ○弁別的証拠［それぞれの概念によって説明される分散部分が，概念間の相関の２乗よりも大きい，仮説モデルと，それより少ない因子を持つモデルとを比較し，その χ^2 値の差を検定する，など］ ○法則的証拠［各構成概念間の因果関係を示すパス係数が有意かどうか］

4.1　r の分布

標本の相関係数：r，母相関係数：ρ とする．

（ｉ）$\rho = 0$ のとき

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{は自由度：} \phi = n-2 \text{ の } t \text{ 分布に従う} \tag{1}$$

（ii）$\rho \neq 0$ のとき

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r} \quad \text{は近似的に正規分布 } N\left(\frac{1}{2}\ln\frac{1+\rho}{1-\rho},\ \frac{1}{n-3}\right) \text{に従う}^{9} \tag{2}$$

[9] 正規分布は $N(\mu, \sigma)$ で表される．

4.2　母相関係数の検定を利用する場合

「因子間の相関係数が１と有意に異なること」という基準を用いる場合，帰無仮説 H_0：$\rho = \rho_0 (\neq 0)$ のとき，対立仮説 H_1：$\rho \neq \rho_0$ となる．有意水準は通常，5%（0.05）あるいは1%（0.01）が用いられる．両側検定なので5%のときは0.025を，1%のときは0.005を計算で用いることになる．以下，簡略化のために，%点をあらわす文字 α は式中において，厳密には $0.01 \times \frac{\alpha}{2}$ を表すものとする．

$$\eta = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} \tag{3}$$

とおけば，式 (2) より

$$U = \frac{Z - \eta}{1/\sqrt{n-3}} \quad \text{は近似的に標準正規分布 } N(0, 1) \text{ に従う}^{10} \tag{4}$$

そこで，U の実現値を u_0 とすると $(\eta,\ \rho \rightarrow \eta_0,\ \rho_0)$

$$\mathrm{R} : |u_0| \geq u(\alpha) \tag{5}$$

求めた個々の相関係数が，式 (5) で求めた棄却域：R に入れば，帰無仮説は棄却され，対立仮説が採択される．

弁別妥当性の検証の場合，「因子間の相関係数が１と有意に異なること」を証拠として提出できる．しかし，式 (3) より $\rho_0 = 1$ とはできないため，$\rho_0 = 0.999$ などとして計算し，帰無仮説の棄却を目指すことになる．

ここで，実際の検証力を確かめるために，サンプルサイズ：n（28，103，2503の３通り），母相関係数：ρ（0.99，0.999の２通り），有意水準：α（0.1，10の２通り）の各値を変化させた時に，帰無仮説を棄却できるような r の値を求めたものが表３である．

表３　実験（検定）

n	ρ	η	α	u (境界値)	Z	r
28	0.99	2.646652	0.1	0.0005	2.646752	0.990002
28	0.99	2.646652	10	0.05	2.656652	0.990197
28	0.999	3.800201	0.1	0.0005	3.800301	0.999000
28	0.999	3.800201	10	0.05	3.810201	0.999020
103	0.99	2.646652	0.1	0.0005	2.646702	0.990001
103	0.99	2.646652	10	0.05	2.651652	0.990099
103	0.999	3.800201	0.1	0.0005	3.800251	0.999000
103	0.999	3.800201	10	0.05	3.805201	0.999010
2503	0.99	2.646652	0.1	0.0005	2.646662	0.990000
2503	0.99	2.646652	10	0.05	2.647652	0.990020
2503	0.999	3.800201	0.1	0.0005	3.800211	0.999000
2503	0.999	3.800201	10	0.05	3.801201	0.999002

[10] 標準化は $z = \frac{x - \mu}{\sigma}$ で求められる．

表３からわかることは，n，ρ，αの値をかなり極端な値にしても，標本の相関係数は少なくとも0.99以上の値がなければ，帰無仮説は棄却されてしまうということである．すなわち，標本の相関係数が少なくとも0.99以上でなければ，相関係数は１と有意に異なり，すなわち弁別的証拠が示されたことになってしまう．したがって，この方法では弁別的証拠の検証力がほとんどないと考えられる．

4.3　母相関係数の推定を利用する場合

「母相関係数の99％信頼区間が1.0を含むかどうか」という基準を用いる場合，まずrをZ変換すると

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r} \tag{6}$$

次に4.2と同様に有意水準αを定め，ηの信頼区間を求める

$$\eta = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \tag{3}$$

とおけば，式(2)より

$$U = \frac{Z-\eta}{1/\sqrt{n-3}} \quad \text{は近似的に標準正規分布 } N(0,\,1) \text{ に従う} \tag{4}$$

そこで，

$$\Pr\left\{-u(\alpha) \le \frac{Z-\eta}{1/\sqrt{n-3}} \le u(\alpha)\right\} = 1-2\alpha \tag{7}$$

$$\Leftrightarrow \Pr\left\{Z - \frac{u(\alpha)}{\sqrt{n-3}} \le \eta \le Z + \frac{u(\alpha)}{\sqrt{n-3}}\right\} = 1-2\alpha \tag{8}$$

よって，

$$\left(Z - \frac{u(\alpha)}{\sqrt{n-3}},\, Z + \frac{u(\alpha)}{\sqrt{n-3}}\right) = (\eta_1,\, \eta_2) \tag{9}$$

逆変換によりρの信頼区間を求める．式(9)のとき，

$$\eta_1 \le \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \le \eta_2 \tag{10}$$

$$\Leftrightarrow e^{2\eta_1} \le \frac{1+\rho}{1-\rho} \le e^{2\eta_2} \tag{11}$$

$$\Leftrightarrow \frac{\exp(2\eta_1)-1}{\exp(2\eta_1)+1} \le \rho \le \frac{\exp(2\eta_2)-1}{\exp(2\eta_2)+1} \tag{12}$$

ここで，実際の検証力を確かめるために，サンプルサイズ：n，有意水準：α，サンプルの相関係数：r の各値を変化させた時に，母相関係数：ρ の取りうる値の範囲を求めたものが表４である．

表４　実験（推定）

n	α	TINV	r	η_1	min: ρ	η_2	max: ρ
28	0.1	3.725144	0.99	1.901624	0.956376	3.391681	0.997738
28	0.1	3.725144	0.999	3.055172	0.99557	4.54523	0.999775
28	10	3.725144	0.99	1.901624	0.956376	3.391681	0.997738
28	10	3.725144	0.999	3.055172	0.99557	4.54523	0.999775
103	0.1	3.390491	0.99	2.307603	0.980394	2.985702	0.994912
103	0.1	3.390491	0.999	3.461152	0.998031	4.13925	0.999492
103	10	3.390491	0.99	2.307603	0.980394	2.985702	0.994912
103	10	3.390491	0.999	3.461152	0.998031	4.13925	0.999492
2503	0.1	3.294423	0.99	2.580764	0.988599	2.712541	0.991229
2503	0.1	3.294423	0.999	3.734313	0.998859	3.86609	0.999123
2503	10	3.294423	0.99	2.580764	0.988599	2.712541	0.991229
2503	10	3.294423	0.999	3.734313	0.998859	3.86609	0.999123

（注）TINVはstudentのt分布の逆関数値を表す．また母相関係数ρの最小値は「min: ρ」，最大値が「max: ρ」をそれぞれ表す．

表４からわかることは，n，α，r の値をかなり極端な値にしても母相関係数の取りうる範囲に１は含まれないということである．すなわち，標本の相関係数が非常に高い値であっても母相関係数の信頼区間にほとんど１を含むことはなく，よって弁別的証拠が示されたことになってしまう．したがって，この方法では弁別的証拠の検証力がほとんどないと考えられる．

4.2や4.3の結果から，弁別的証拠の検証には「それぞれの概念によって説明される分散部分が，概念間の相関の２乗よりも大でなければならない」という基準を用いるのがよいと考えられる．

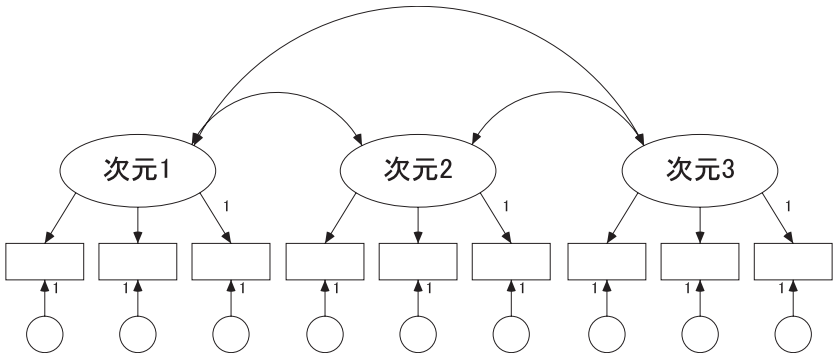
５．法則的証拠に関する検証方法の検討

法則的証拠とは，構成概念が理論や仮説の中で有している役割に応じて，他の構成概念に対し法則的な関連（因果関係や相関関係）を示すという証拠である．具体的には「各構成概念間の因果関係を示すパス係数が有意かどうか（趙, 兪 2004）」などの基準を用いるべきであるとされている．しかしながら，法則的証拠は構成概念妥当性の中で言及されないことが多い．その大きな理由の一つは，尺度開発など一つの構成概念に対して多くの項目を用いた研究における法則的証拠の検証では，モデルが複雑になりやすくモデル全体の適合度指標が満足な値にならないことがほとんどであるということがある．つまり，法則的証拠を示すためには対象となる構成概念を中心とした近接構成概念間の因果関係モデルが必要となる．しかし同時に，尺度開発などでは因子分析モデルが用いられている．これら２つのモデルを単純に重ね合わせる（つまり近接構成概念間の因果関係モデルにそのまま因子分析モデルを導入する）と非常に複雑なモデルができてしまう．また，尺度開発と法則的証拠の検証を分けてしまい，それぞれ同じ構

成概念を表すとされている別の項目で考えるというやり方もあるが，このやり方では両者が本当に同じ構成概念を測定しているという保証はまったくない．そこで本稿ではこの問題を解決するために，２次因子構造型の多重指標モデルとItem Parcelingの導入を提案したい．

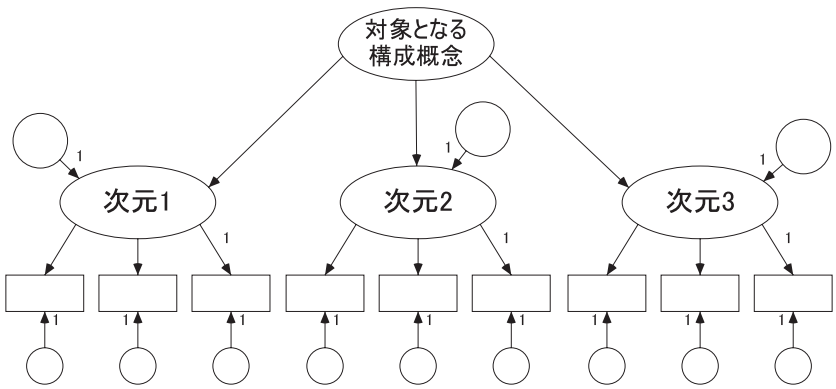
尺度開発の多くは特にモデルが明示されていない場合，因子分析モデルを用いて行われている．そこでは対象となる構成概念のいくつかの次元を探索的因子分析から見出し，それらを外生変数とし，尺度項目をいずれかの構成概念の測定項目とするモデルが想定される（図１）．

図１　因子分析モデル



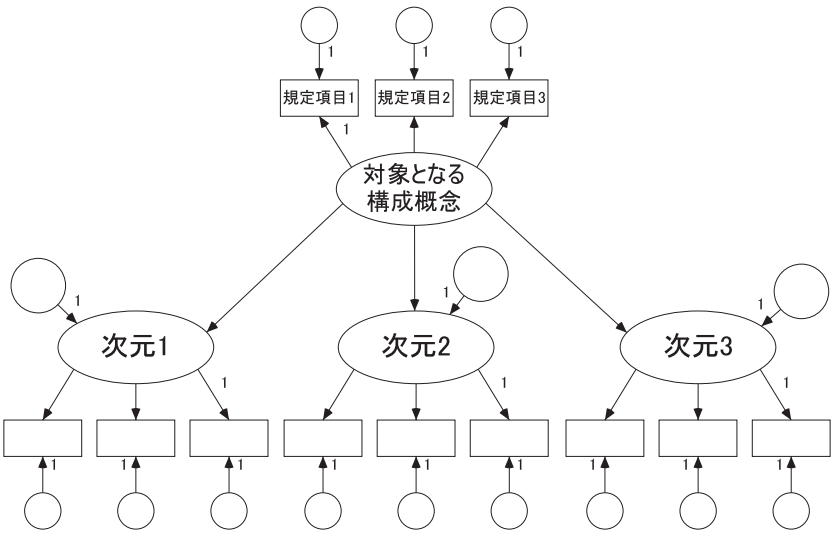
しかし，見出される次元が対象とする構成概念の様々な側面を表しているとするならば，対象構成概念を２次因子とする２次の因子構造（図２）を想定する方が，理論的に筋が通っていると考えられる．ただし，豊田（2005）は２次因子分析を用いた方法の問題点として，下位指標の重みづき合計で算出される２次因子部分は，内的整合性が高すぎても低すぎても不都合であるという点を指摘している．内的整合性が高すぎれば下位指標の意味が薄まるし，内的整合性が低すぎれば２次因子の意味が薄まることになるからである．

図２　２次因子分析モデル



2001年より日経BPコンサルティングが行っている「ブランド・ジャパン」というブランドのランキング・プロジェクトがある[11]．ブランド・ジャパンでは，消費者が評価するBtoC指標と，ビジネス･パーソンが評価するBtoB指標があり，それぞれ異なるモデルが共分散構造分析によって分析されている．BtoC指標は２次因子分析モデル（図２のようなタイプ）が採用され，２次因子として「対象となる構成概念＝ブランドの総合指標」，１次因子として「フレンドリ」「イノベーティブ」「アウトスタンディング」「コンビニエント」の４つの下位次元が想定されている．一方，BtoB指標では２次因子構造型の多重指標モデルが採用されている．BtoCと近い形をしているが，２次因子に測定項目が付けられている点で大きく異なる（図３）．

図３　２次因子構造型の多重指標モデル



この２次因子構造型の多重指標モデルの特長は，対象となる構成概念に直接付けられた項目によって，構成概念の内容を規定できることにある．言い換えると２次因子型の多重指標モデルでは，対象となる構成概念を２通りで示していることになる．つまり一方が上半分の測定方程式の部分（対象となる構成概念を規定項目１～３で測定する部分）であり，もう一方が下半分の２次の因子分析モデルの部分（図２の部分）である．そして２次因子構造型の多重指標モデルのこの性質に着目し，構成概念の因果関係モデルとの関連づけを行うことができる．具体的に図で説明すると，対象となる構成概念とその近接構成概念A，B，Cの因果関係を，例えば図４のように設定したとする．このときに，対象となる構成概念を測定方程式モデルで表わしたのが図４であり，２次因子型モデルで表わしたのが図５である．図４と図５で用いられている「対象となる構成概念」は，図３によって同じものであることが担保されている．つまり，図３によって図４と図５が関連づけられているのである．

ただし，実際には図５のようにモデル化すると，次元が多い場合など観測指標の数が莫大に

[11] 詳しくは，http://consult.nikkeibp.co.jp/consult/br/bj2008/index.htmlを参照のこと．

なり満足な適合度を得ることが非常に困難になる．ゆえに，Item Parceling[12] などにより尺度項目を合成するなどして，モデルが肥大化するのを防ぐ必要がある（図６）．

図４　構成概念間の因果関係モデル

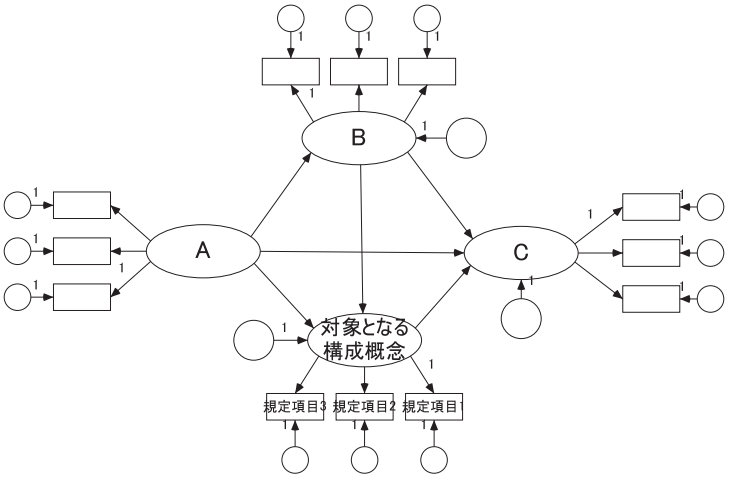
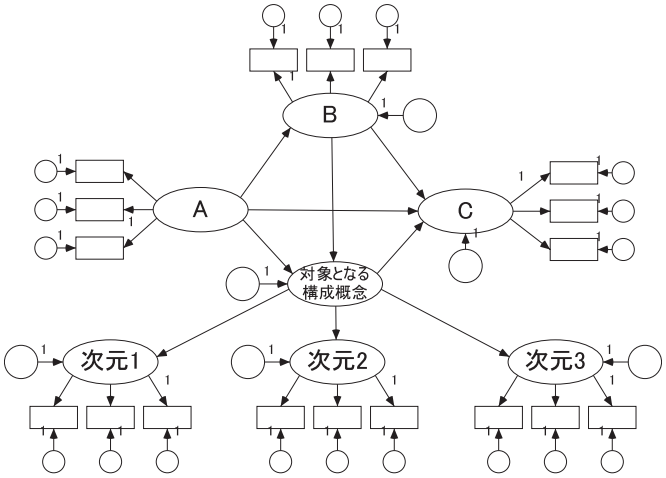


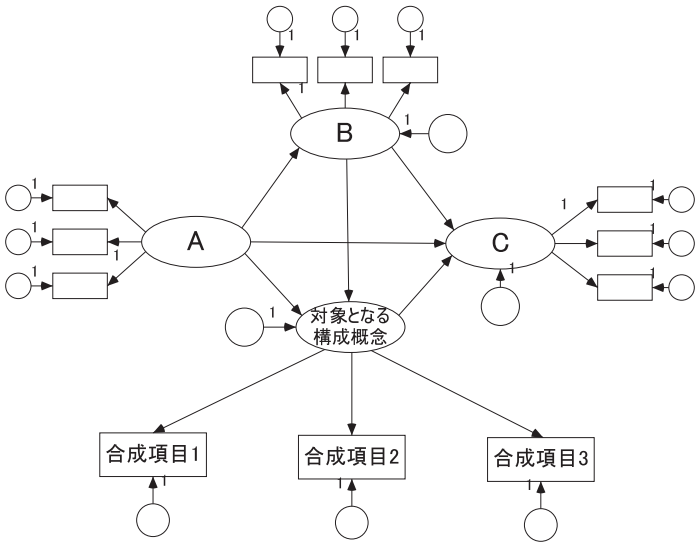
図５　構成概念間の因果関係モデル（２次の因子構造を代入）



[12] 項目得点の和や平均値を利用して，複数の項目をまとめた新しい変数を作成し，分析に利用すること．一定の条件を満たすことで，「全項目を単純に足し合わせて観測変数を減らす」ための応急措置に止まらず，「適切な項目の組み合わせによって分析の精度を高める」ための積極的な分析技法となる（室橋 2004）．問題はいくつかあるが，特に問題となるのはどの項目でパーセルを作るかということである．Bandalos（2002）は，多次元である項目群，あるいは因子構造がわからない項目群に対してParcelingを用いることは勧められない，としている．星野・岡田・前田（2005）では，Parcelingを行うのなら，合計した項目の一次元性を因子分析などでテストする必要があるとしているし，室橋（2004）でも，Parcelingする項目間に多次元的な関係がないかどうかを確認する必要がある，としている．

したがって，図３の形で尺度開発を進めると，構成概念の因果関係モデルに関連づけられた尺度が作られることになる．構成概念間の因果関係モデルのなかで尺度を位置づけることができると，少なくとも３つの点で利点がある．それは，対象となる構成概念の構造を捉えることで評価が改善しやすくなるという点，構成概念妥当性を検討する際，法則的証拠を併せて吟味できるためより妥当性の高い尺度の開発が可能になると考えられる点，そして他の構成概念に関する測定項目を共に問うことで，概念の区別がなされやすくなり，対象となる構成概念をより明確に捉えることができる点である．

図６　構成概念間の因果関係モデル（次元の合成項目を代入）



６．おわりに

本稿では構成概念妥当性の概念の変遷と各側面の証拠の検証方法を具体的に示し，整理した．さらに弁別的証拠と法則的証拠に関して取り上げ，それぞれの問題点の指摘と提案を行った．弁別的証拠に関しては，消費者行動分野においてこれまで最も広く用いられてきた，母相関係数の検定や推定を利用した検証方法が検証力を持たないことが明らかになった．また連続的なχ^2検定による方法も扱う変数の数が増えると現実的に用いるのが厳しくなることから，最終的に「それぞれの概念によって説明される分散部分が，概念間の相関の２乗よりも大でなければならない」という基準を用いるべきであるという提案を行った．法則的証拠に関しては，尺度開発など一つの構成概念に対して多くの項目が尺度として用いられるとき，構成概念間の因果関係モデルと対象構成概念の因子分析モデルとを重ね合わせると，モデルが複雑になるために全体として満足な適合度指標を得ることができなくなり，法則的証拠が示されない場合が多いことが指摘された．しかし，因果関係モデルと因子分析モデルを切り離して議論するのは，想定する構成概念が同じであるという保証が得られず，問題があることがわかる．したがって，これら２つのタイプのモデルを用いるときには，２次因子構造型の多重指標モデルとItem

Parcelingを使用すべきであることが提案された.

構成概念妥当性の検証は心理変数を用いるような定量調査において非常に重要であるにもかかわらず，構成概念妥当性の意味を考えることなしに，単に決まった手順を盲目的に繰り返しているだけの研究が少なくない．研究者が恣意的に設定した基準値を上回れば信頼性が十分であるとか，収束妥当性が高いとかいった議論では，本当の意味での構成概念妥当性の検証には当たらない．結果的に同じ手順と基準を用いることになったにせよ，構成概念妥当性について十分に考慮した上での結果ならば，十分に吟味することによって自ずと妥当性は高くなり，一見同じに見える結論であってもその後に続く研究への示唆の大きさはまったく異なるものになるのである.

謝　辞

阿部周造先生からは，大学学部，大学院修士課程・博士課程と７年半に渡りゼミナールにおいて直接ご指導を頂きました．専門領域の全体に広がる豊富な知識と真理を見つめる深い洞察力はもちろんのこと，研究に向かう謙虚でひたむきな姿からは非常に多くの感銘と教訓を得ることができました．この場をおかりして感謝申し上げます.

参 考 文 献

阿部周造（1981），「消費者情報処理の経験的研究」，『マーケティング・ジャーナル』，1（3），pp.12-22.
阿部周造（1984），「消費者情報処理理論」，中西正雄（編），『消費者行動分析のニュー・フロンティア　―多属性分析を中心に―』，誠文堂新光社，pp.119-163.
阿部周造（1987），「構成概念妥当性とLISREL」，奥田和彦，阿部周造（編），『マーケティング理論と測定―LISRELの適用―』，中央経済社，pp.27-46.
阿部周造（2002），「仮想的消費データに基づく満足研究の妥当性」，『明大商学論叢』，84（1），pp.1-10.
川上智子（2005），『顧客志向の新製品開発　―マーケティングと技術のインタフェイス―』，有斐閣.
久保田進彦（2006），「リレーションシップ・マーケティングのための多次元的コミットメントモデル」，『流通研究』，9（1），pp.59-85.
清水裕子（2005），「測定における妥当性の理解のために　―言語テストの基本概念―」，『立命館言語文化研究』，16（4），pp.241-254.
鈴木宏哉（2005），「サッカーのゲームパフォーマンス尺度と因果構造」，博士論文（筑波大学）.
趙顕哲，兪在沅（2004），「顧客の役割認識および知覚されたリスクとサービス接点満足との関係」，阿部周造，新倉貴士（編），『消費者行動研究の新展開』，千倉書房，pp.59-74.
豊田秀樹（2005），「「ブランド・ジャパン」分析方法の解説」，日経BPコンサルティング　（http://consult.nikkeibp.co.jp/consult/br/bj2005/rep04.html）.
中村健太郎（2003），「一般化可能性理論，多変量一般化可能性理論」，豊田秀樹（編），『共分散構造分析［技術編］』，朝倉書店，pp.71-78，79-89.
畑井佐織（2004），「消費者とブランドの関係の構造と測定尺度の開発」，『消費者行動研究』，10（1・2），pp.17-41.
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.
星野崇宏，岡田謙介，前田忠彦（2005），「構造方程式モデリングにおける適合度指標とモデル改善について　―展望とシミュレーション研究による新たな知見―」，『行動計量学』，32（2），pp.209-235.
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.
村上宣寛（2006），『心理尺度のつくり方』，北大路書房.
室橋弘人（2004），「Item Parceling の効用と限界」，第68回　日本心理学会大会ワークショップ報告資料.
吉田富士雄（1994），「心理尺度の信頼性と妥当性　―尺度が備えるべき基本的条件―」，『心理尺度ファイ

ル　―人間と社会を測る―』，堀洋道，山本真理子，松井豊（編），垣内出版，621–35.
American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1985), *Standards for Educational and Psychological Tests*, American Psychological Association: Washington.
American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999), *Standards for Educational and Psychological Testing*, American Psychological Association: Washington.
American Psychological Association (1954), Technical Recommendations for Psychological Tests and Diagnostic Techniques, *Psychological Bulletin*, 51 (2), 1–38.
American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1966), *Standards for Educational and Psychological Tests and Manuals*, American Psychological Association: Washington.
American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1974), *Standards for Educational and Psychological Tests*, American Psychological Association: Washington.
Bandalos, Deborah L. (2002), "The Effects of Item Parceling on Goodness-of-Fit and Parameter Estimate Bias in Structural Equation Modeling," *Structural Equation Modeling*, 9 (1), 78–102.
Hively, Wells, Ⅱ, Harry L. Patterson and Sara H. Page (1968), "A UNIVERSE- DEFINED System of Arithmetic Achievement Tests," *Journal of Educational Measurement*, 5 (4), 275–90.
Messick, Samuel (1989), "Validity," in *Educational Measurement*, ed. Robert L. Linn, New York: Macmillan Publishing Company, 13–104.［邦訳：池田央，藤田恵璽，柳井晴夫，繁桝算男（編・訳）(1992)，『教育測定学』，C.S.L.学習評価研究所，147–209.］
Messick, Samuel (1995), "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning," *American Psychologist*, 50 (9), 741–49.
Osburn, H. G. (1968), "Item Sampling for Achievement Testing," *Educational and Psychological Measurement*, 28 (1), 95–104.
Popham, W. James and IOX Assessment Associates (1997), "Consequential Validity: Right Concern-Wrong Concept," *Educational Measurement: Issues and Practice*, 16 (2), 9–13.

〔なかむら　あきと　福島大学経済経営学類准教授〕
〔2009年５月１日受理〕