大学評価・学位研究　第18号　平成29年３月（研究ノート・資料）
［独立行政法人大学改革支援・学位授与機構］

自己評価力向上支援のための評価指標設定に関するチェックリストの開発

Checklists for setting indicators to enhance the self-assessment capability of universities

渋井　進，浅野　茂，橋本　貴充，小野　宏，
小野　達也，山﨑　その，田中　弥生
SHIBUI Susumu, ASANO Shigeru, HASHIMOTO Takamitsu, ONO Hiromu
ONO Tatsuya, YAMASAKI Sono, TANAKA Yayoi

Research on Academic Degrees and University Evaluation, No. 18 (March, 2017) [the essay/material]
National Institution for Academic Degrees and Quality Enhancement of Higher Education

自己評価力向上支援のための評価指標設定に関するチェックリストの開発

渋井　進*，浅野　茂**，橋本　貴充***，小野　宏****
小野　達也*****，山﨑　その******，田中　弥生*******

要　旨

今日，政策的な流れとして，成果を定量的に示すことが求められ，公的資金給付の要件として指標の設定とそれを用いた実績報告が求められるようになっている。こうした状況に鑑み，大学評価，IR 室に対して指標のデザインとデータ分析の業務の重要性が増している。他方で，大学評価室，IR 室などにおける主たる業務として，目的・計画の達成状況を測定する指標をデザインするが，その際の問題として，指標が内容を適切に捉えているか，評価機関等において指標を用いて評価する側の評価者を十分に説得できる内容であるかという妥当性の問題がある。この問題解決を支援するために，簡便なチェックリストをその解説とともに開発した。本論文では，最初に現場における大学評価と指標の問題についての具体例を示しながら問題提起を行う。次に，その解決へ向け，指標設定の信頼性・妥当性について先行研究として代表的な政策評価及び心理測定における指標研究のレビューを行う。それを踏まえ，妥当性の意味概念の相違や大学への適用可能性を検討することで，チェックリストを開発した。

キーワード

評価指標，妥当性，信頼性，政策評価，心理測定

1．はじめに

1.1　大学評価と指標の問題

一般に，評価をするにあたり，その対象となる目的・目標や，基準・観点等に対応した根拠となるデータを収集し，いかに指標を設定するかは，評価における根本的かつ重要な問題といえる。特に，大学評価室，Institutional Research（IR）室等から見ると，国立大学法人評価の年度実績評価や中期目標期間の評価，競争的資金や概算要求の採択後の評価など，公的資金配分の要件としてアカウンタビリティーを遂行すべく，定量的な指標設定とその測定による評価改善サイクルの確立が，ここ数年強く求められていることもあり，評価指標の重要性に関わる意識は高まりつつあるといえる。

具体例として，国立大学法人評価では，第 3 期中期目標期間における運営費交付金の配分方法について，「人材育成や地域課題を解決する取組などを通じて地域に貢献する取組とともに，専門分野の特性に配慮しつつ，強み・特色のある分野で世界ないし全国的な教育研究を推進する」，「専門分野の特性に配慮しつつ，強み・特色のある分野で地域というより世界ないし全国的な教育研究を

*　独立行政法人　大学改革支援・学位授与機構　研究開発部　准教授
**　山形大学　学術研究院　教授
***　独立行政法人　大学入試センター　研究開発部　助教
****　関西学院大学　企画室　総合参事
*****　鳥取大学　地域学部　教授
******　京都外国語大学　総合企画室　次長
*******　独立行政法人　大学改革支援・学位授与機構　研究開発部　教授

推進する」,「卓越した成果を創出している海外大学と伍して，全学的に世界で卓越した教育研究，社会実装を推進する」という三つの大学の類型に基づいた重点支援の枠組みを示していることが，社会的に大きな関心を集めているところである。その中の予算配分の具体的な決定方法において「各国立大学法人が，取組構想の内容に応じて，中期目標期間を見通した取組の成果を検証するため，原則として測定可能な評価指標（KPI）を独自に設定する」,「評価指標については，その妥当性を裏付けることができるよう，各国立大学法人において比較すべき指標（ベンチマーク）や客観的な根拠を用意する。」（第３期中期目標期間における国立大運営費交付金の在り方に関する検討会，2015）とあるように，大学自らが客観的に測定可能な指標を設定することが求められており，評価指標の設定に関する関心は高まっている。

また，競争的資金の評価の例では,「スーパーグローバル大学創成支援」事業や,「大学教育再生加速プログラム」,「地（知）の拠点大学による地方創生推進事業」などの文部科学省の補助金申請時にも，申請要件として大学に共通に既に設定されている指標や，年度ごとに予測を記入する成果指標がある一方で，大学独自での指標設定が求められている。すでに設定されている指標についても，例えば,「地（知）の拠点大学による地方創生推進事業」（COC＋）において,「数値目標については，目標値の妥当性や設定した根拠を説明してください。（文部科学省，2015）」とあるように，数値の妥当性について，大学自らが判断して値を設定することが求められており，これらに対処するためにも，大学の指標に対する関心は高まっているといえる。

このように，大学が指標の妥当性を考慮し，自ら設定することが求められているが，大学の評価担当者レベルでは，実際にどのような指針を持って行えば良いか悩ましい現状である。例として，国立大学における第１期中期目標期間の評価の際には，達成状況報告書における中期目標・計画に対応するアウトカムが想定されておらず，データが体系的に収集されていないという問題があった（渋井・面高，2009）。第１期の評価において，アウトカムを示す指標収集が困難であった理由には，目的および，計画の立て方の問題もあったといえる。こうした問題の解決に寄与すべく,田中（2009）は，評価の視点を取り入れて，目的と計画の立て方を見直す方法を Evaluability Assessment（評価可能性のアセスメント，EA）として提案した。その後，田中らは，大学評価・学位授与機構の EA 研究会を通じて，支援ツールの提供と普及を図って来た（大学評価・学位授与機構，2013a）。

他方，国立大学法人第２期中期目標・中期計画期間では，先のような大学への文部科学省の改革補助金による支援がみられ，また，IR の整備に伴い根拠資料・データの継続的な収集とチェックに対する意識は高まりつつある。中でも，データの根拠となる指標の問題については，ますます関心が高まっている。だが，設定する指標の妥当性については，大学の評価担当者が，ワーキンググループ等を開催し，その都度検討していることが多く，その負担は大きく，何らかの支援が必要とされているといえる。

そこで，大学以外の組織に目を向けてみると国際協力機構（JICA）の試みがある。JICA は国際援助協力活動の効果を測定するために指標の選定やデザインに注力してきており，評価指標選定のための１つのツールとして無償資金協力事業における開発課題別の標準的指標例（ガイドライン）を作成した（JICA，2013）。このガイドラインは，JICA に蓄積された指標データを参照して構築されたもので，途上国における開発課題に関し，案件形成，事前評価段階における定量的効果に係る指標の設定時にレファレンスとして用いられることを目的としており，行政改革推進会議でも行政改革のための優良事例として注目されている。このように過去の事例集を作成し，参照することは，１つのツールとして有効な方法と考えられる。過去のデータをレファレンスとして用いるという点では，大学では経験的に，認証評価において，過去の類似した他大学の根拠資料・データ等を参照するということが行われている。しかし，機関別認証評価のように基準・観点が定まった評価では有効であるが，国立大学法人評価のように，それぞれの大学の事情を踏まえて大学固有の目的・目標が設定されている場合や，競争的資金の申請書においてもそれぞれの大学の現状を把握した上での事業目的に応じた指標設定が必要なため，その多様性から事例集にも限界があると考えられる。

1.2　指標にかかる大学現場の状況

前節で述べた様に，指標設定の必要性に対する大学の意識は高まっている。本節では，大学における評価指標について，それが求められるようになっている背景，具体的な設定にまつわるプロセス，設定の難しさ等についての現状を反映する例として，指標設定について先進的と考えられる二大学（関西学院大学，京都外国語大学）を選び予備的にヒアリングを行った。その結果，以下のような課題が得られた。

○関西学院大学へのヒアリング結果

・インプット，アクティビティ，アウトプット，アウトカムといった指標の類型を，どのように使い分ければいいのか。
・目的体系図の下位と上位でどの種類の指標を設定するのが適切か。
・指標は一つの施策に対して一つであるべきか，あるいは複数が良いのか。
・定量的な指標と定性的な指標をどのように使い分けるのか。
・新たな指標を開発する場合に，注意すべきことは何か。
・データを入手するのに労力やコストがかかる場合にどう考えるか。
・直接的に測定できない場合に間接的な代替の指標をどうやって開発するか。
・自分たちが現在設定しているさまざまな指標が，測るべき対象・内容に対して適切かをどう判断するのか。
・一般的に評価指標の妥当性が，どのような条件・要素によって担保されるのか，指針となるような理論的な裏付けはないか。

○京都外国語大学へのヒアリング結果

・近年の文部科学省の補助事業の申請書において事前に設定され，大学に要求されている指標では，外形的な大学改革に着手しているかの評価が中心となっており，確かな成果をあげているかどうかは，大学がそれぞれの目的に沿って独自の指標を設定する必要がある。
・これまでの自己評価は，認証評価受審の一環として行っていた部分が大きく，大学自身で５カ年計画を策定することとなり，それにあわせて目標達成を意識した指標の設定が望まれている。
・各計画の目的に対する指標は，執行部が経営的な視点で考え，各年度の実施事業に対する指標は主担当者の学科や部署が現場の視点で考えるというように異なる視点から設定したため，なぜその指標が必要なのかは各々の立場からは説明できるが，指標間にどのような関係があり，それを辿っていけば目的に行きつけるという因果関係は説明できない。
・適切な指標の設定は，計画の策定，評価のどちらにとっても重要な事項であるが，計画全体の改善を図るPDCAサイクルとは別の視点でチェックすることが必要である。

以上のように，両者のうち一方は比較的規模の大きな総合大学であり，もう一方は比較的小規模の単科系の大学であるが，目標・計画を達成するために，妥当な指標をいかに設定するか，という悩みでは共通する点が多いことがわかる。

関西学院大学において挙げられた，「新たな指標を開発する場合に，注意すべきことは何か」「一般的に評価指標の妥当性が，どのような条件・要素によって担保されるのか，指針となるような理論的な裏付けはないか」に見られるように，指標の妥当性の判断に伴う困難さの問題が，現場においてみられる。本論文は，ヒアリング結果の全てにこたえるものではないが，課題解決の一助として，妥当性について扱った。なお，ここで扱う妥当性とは，大学が抱いている指標設定の課題群を，妥当性の課題であると総称するものであり，その概念整理を行った。

2．目的

本研究は，指標の妥当性や信頼性を確認するためのチェックリストを開発することを目的とする。

先の２つの例に挙げた様に，文部科学省の改革補助金を取得し，先端的な取組を行っている私立大学においても，補助金申請時の評価や中間評価に対応する中で，個別の計画における目標達成度の評価においては標準的な指標はなく，大学が指標を独自で設定する必要があることから，その際の妥当性についての判断が必要となり，そのために，理論的裏付けを持った妥当性の判断基準の必要性が生じている。

したがって，大学現場のニーズに応えるものを開発する必要があると考えた。そこで想起されたのが，指標設定の際の目安になるもの，すなわち

チェックリストであった。しかしながら，安易なチェックリストはすぐにその有用性を失うだけではなく，大学現場を混乱させ負荷をかけてしまう可能性もある。そこで，理論的な基礎や背景を抑えたうえで，目安となるべきチェックリストを作成する必要があると考えた。このチェックリストは直接的に現場の課題解決の方法を示すものではないが，妥当性について学術的な背景をもとに概念整理をしたものである。現場の多様性により課題解決には，各大学の目的や置かれた状況等に沿ってチェックリストの細分化が必要となると思われるが，ここでは一つの目安としてのチェックリストの提案を目的とする。

3．指標の妥当性に関する先行研究のレビュー

3.1　大学評価の文脈における指標の「妥当性」

大学評価の妥当性については，特に我が国においてはその歴史の短いこともあり，目標達成度型の評価における指標設定の妥当性という問題そのものについ扱われることはなかった。大学評価の一部として，教育評価の中でも，特に教育測定の分野においてテストの妥当性や，授業評価の妥当性について扱った研究もあるが（大塚，2007），基本的には後述する心理測定と関連し，信頼性と妥当性について検討した内容に留まっている。

以上の様に，大学評価そのものを対象として，指標の妥当性そのものを検討した学術文献を見出すことはできなかった。ただし，指標設定を考慮するために提供されている情報という点で，関連するいくつかの文献を以下に挙げることとする。

評価機関側から提供されている資料の例として，大学評価・学位授与機構では，認証評価に関しては自己評価実施要項（大学評価・学位授与機構，2014）における「観点に対する関係法令及び分析する際の留意点，根拠資料・データ等例」，国立大学法人評価に関しては実績報告書作成要領（大学評価・学位授与機構，2013b）の「『教育の水準』，『研究の水準』の観点ごとの分析に当たっての留意点等」が挙げられる。これらは評価基準や観点についての解説および，指標例が示されており，どのような指標・エビデンスを記述するかを検討する上で，大学にとって有益な資料である。しかしながら，基準，観点に沿って評価機関から示されているものであり，国立大学法人評価における中期目標の達成度の評価についての指標導出に関したマニュアル等はなく，妥当性の判断についての言及がなされているものでもない。

一方，大学の側での指標作成に関連し，大学評価コンソーシアムにおいて作成された「データ収集作業のガイドライン」（大学評価コンソーシアム，2013）がある。これは，データ収集の課題と改善のための手がかりについてまとめたものである。これも指標の妥当性の判断について直接的に述べてはいないが，現場での指標設定において，データが実際に学内で収集可能かどうか，という点が考慮されることは多く，その点では妥当性と関係しているともいえる。

3.2　複数分野における先行研究

以上の様に，大学評価において指標の妥当性そのものをどう扱うかについて，定まった議論は無い。そこで，本節ではいくつか他の参考となる研究分野において妥当性がどのように扱われてきたかについてレビューを行う。具体的には，指標の信頼性や妥当性について先行研究がある政策評価と心理測定[1]という 2 つの分野における妥当性の扱いについてレビューする。

妥当性について，古くは Campbell and Stanley（1963）において論じられた内的妥当性と外的妥当性という 2 つの分類が挙げられる。これは，実験計画法の文脈で論じられた概念であり，内的妥当性とは，実験によって得られた結果が実験手続きによる影響に基づくものかという，その実験内部での因果推論に関するものであった。一方，外的妥当性とは得られた結果がどれだけ一般化出来るか，外的にどの程度通用するかという，一般化可能性に関するものであった。彼らの論文では，

[1] 政策評価は行政府機関を対象にした評価の考え方を示すものであるが，教育・研究という公的機能を果たし，非市場における活動に従事するという点で大学と行政府機関は共通する点が多い。中でも，国立大学法人法は，独立行政法人法通則法を参照して作成されており共通する点が多い。また，心理測定法は教育心理学とその理論的，技術的なベースを共有している。

それぞれの妥当性を脅かす要因として，内的妥当性について8つ，外的妥当性について4つを挙げ，解説をしている。

Campbell and Stanley（1963）はその後も政策評価，心理測定，医学等の幅広い分野において，妥当性を検討する際に触れられることが多く（Chen et al., 2011；成田，1986など），妥当性について最初に問題提起した研究といえるだろう。その後，分野ごとの特性を踏まえた妥当性の文脈等を考慮しつつ，妥当性概念の検討と細分化がなされて来たといえる。

一部，教育評価の分野においては，教育測定における信頼性・妥当性（後述する心理測定における信頼性・妥当性概念とほぼ同義）をベースとした実証主義的なアプローチに対し，評価者と被評価者の関係性を重視する構成主義的立場から，評価のパラダイムシフトの必要性を論じる立場もある（Guba & Lincoln, 1994；北川，2008）。しかし，これは妥当性を扱ったのではなく，より広範に評価の在り方を論じていると解釈されるため，ここでは割愛する。

3.2.1　政策評価における妥当性

政策評価（ここでは公共政策を評価するための様々な方法・制度を総称して政策評価という）においては記録・観察・調査などによって得られる数・量が広く用いられており，それらの数字には当然のこととして妥当性が求められる。ここでは政策評価の分野における妥当性を巡る議論を紹介する。

3.2.1.1　政策評価と大学評価

政策評価の理論・実践を①事前の政策分析（主として費用対効果の観点から評価），②事後のプログラム評価（個々のプログラムを多角的に掘り下げて評価），③事後の業績測定（組織･機構の取り組みを網羅的かつ定常的に評価）という3系譜に分類するとすれば，個々の大学の評価は③の事後評価に相当する。このアプローチ（以下では業績測定型評価と呼ぶ）は欧米ではPerformance Measurementと呼ばれ，1980～90年代以降の公共部門改革の大潮流（New Public Management）における基本ツールとして世界各国で急速に普及したもので，中央・地方政府の政策・施策・事業（これらを総称して以下ではプログラムと呼ぶ）の集合を対象に評価指標と目標値を設定，定期的に達成度評価を行ってPDCAサイクルを回すというのが典型的である。日本でも三重県が1996年度に本格的に導入したのを皮切りに多くの自治体で導入が進み，中央の府省で2001年に導入された政策評価制度でも業績測定型評価の比重は大きい。

個々の大学における評価，国立大学法人評価の年度実績評価や最終評価，競争的資金や概算要求の採択後の評価などは，私学を含め公共的な事業・サービスの評価であるとすれば，それは政策評価に含まれ，その主たる方法は業績測定型評価であるが，大学評価（あるいは学校評価）と一般的な行政プログラムを対象とする政策評価という2つの分野間の連携は，方法開発と実務の両面において乏しいのが現状であろう。なお，認証評価は定期的に行う事後的な評価ではあるが，教育研究活動の質の保証，改善，アカウンタビリティーを目的としている点から，業績測定型評価とは異なるといえる。

3.2.1.2　業績測定型評価における評価指標

・評価指標の種類とプログラムのロジック

業績測定型評価の核心はプログラムの実施結果（特に成果）や効率を評価指標によって測り，その目標達成状況を明らかにする過程にある。評価指標（群）の設定にあたっては，プログラムの最終目的の達成に至るまでの「インプット（予算などの投入）→プロセス（過程）→アウトプット（提供される財・サービス）→直接的アウトカム（直ちに発現する成果）→中間的アウトカム→最終的アウトカム」というロジックを踏まえること，即ち最終成果実現までのどの論理的段階が測定対象なのか明らかにすることがしばしば要請される（その作業としてプログラムのロジックを図示するロジックモデルが描かれる）。

段階として最終的アウトカムが最重要であることは言うまでもないが，評価指標としては，外部要因（他のプログラムなど）の影響や測定費用・所要時間など困難を伴う場合が多い。成果と並ぶ重要概念である効率は，インプットとアウトプット・アウトカムの比として把握するのが基本だが，公共部門においてはプログラムの費用の正確な把握が容易でない（例えば人件費の把握や予算単位との関係）など，こちらも課題は多い。

表１　評価指標が満たすべき条件ー例１

区分		基　準	説　明
個別指標の基準	1–1妥当性	Validity	計測すべき事象を計測
	1–2信頼性	Reliability	正確に計測
	1–3理解可能性	Understandability	意味が明確で誤解しにくい
	1–4タイムリー性	Timeliness	有用なタイミングで入手可能
	1–5目的との適合性	Relevance to the objectives	目的や成果を適切に反映
	1–6施策の影響の大きさ	Program influence	計測事象への施策の影響度
	1–7計測可能性	Feasibility of collecting data	データ収集が可能
	1–8データの収集費用	Cost of collecting data	費用の大きさ
	1–9操作可能性	Manipulability	計測値の操作可能性が小さい
	1–10意思決定への有用性	Usefulness for decision-making	意思決定に有益な知見を提供
指標群の基準	2–1包括性	Comprehensiveness	重要な側面を漏らさずカバー
	2–2非重複性	Nonredundancy	重複なく異なる側面を計測
	2–3データの収集費用	Cost of collecting data	総費用の大きさ
	2–4反抗的行動への耐性	Resistance to perverse behavior	意図に反する行動を誘発せず

田中（2014）をもとに筆者が作成。

表２　評価指標が満たすべき条件ー例２

基準	説明
1 Results oriented	アウトカムに焦点
2 Relevant	目的との関係が論理的かつ直接
3 Responsive	パフォーマンス水準の変化を反映
4 Valid	把握すべき情報を把握
5 Reliable	正確でぶれない情報
6 Cost-effective	データ収集費用が過大でない
7 Useful	意思決定者に有益な情報を提供
8 Accessible	定期的に情報が得られる
9 Comparable	時系列比較が可能
10 Compatible	既存の財務・業務システムに適合
11 Clear	様々な立場の人が理解できる
12 Affordable	予算内で運用できる

ASPA（2000）による（筆者訳）。

・評価指標が満たすべき妥当性及びその他の条件

社会科学の測定に求められる２大条件であるといってよいであろう妥当性・信頼性は，政策評価においても当然求められる。ロジックに基づく指標設定は妥当性を担保する上で本来欠かせない作業であるといえる。業績測定型評価に関するテキストやマニュアルの類は欧米で少なからず刊行されているが，そこではしばしば妥当性・信頼性に他の条件を加えて評価指標が満たすべき条件とされる。評価にとって重要な妥当性は，多義的な概念・用語であり，これについて検討することが求められる。田中（2014）は英語圏で幅広く参照される代表的な文献である Hatry（1999, 2006）及び Ammons（1995）における整理を統合して表１のような基準を提示している。なお，表中の基準には互いに重複する部分もある。

また，米国行政学会（American Society for Public Administration; ASPA）が作成した業績測定型評価のマニュアルも広く利用されており，表２の条件が掲げられている。この中の条件１，９，10は表１に該当するものがない。

3.2.1.3　プログラム評価が求める妥当性等

プログラム評価においては，インパクト（他の要因を除去した正味の成果）を統計解析などにより定量的に明らかにするという観点から，（シンプルな指標による測定を旨とする）業績測定の分野とは異なる流儀で妥当性・信頼性を吟味する。例えばプログラム評価のための評価デザインと統計解析法を統合し体系的に述べたテキストにおいてLangbein（2012）は，妥当性･信頼性を①internal validity, ②external validity, ③measurement validity & reliability，④statistical validity という 4 つに分類している。業績測定における妥当性は①と③の一部，信頼性は②と③の一部に概ね相当する。④は変数間の関係に関する妥当性であり，通常の業績測定には該当しない。

3.2.1.4　目標管理・達成度評価の妥当性

業績測定型評価において指標の妥当性が問題となるのは，実は指標の設定時に限った話ではない。多くの場合，評価指標には目標値が設定され，定期的に実績値の目標達成度を把握する目標管理が行われることとなる。実績値と目標値の比較の妥当性について詳しく述べることは本稿の紙幅では適わないが，妥当性を吟味すべき場面の例として①目標値の設定根拠や性格は妥当か，②達成度の比較は妥当か（例－フロー指標とストック指標は直接比較できない），③達成度の計算は妥当か（例－ストック指標の実績値を目標値で除すと多くの場合意味が曖昧）などを挙げておきたい。

3.2.2　心理測定における妥当性

心理測定において，妥当性とは「研究者によって測定されるデータが，その目的にどれだけかなっているか，特にその概念的な面における適切さの程度」（大津，2011）を意味する。心理学では，心という目に見えないものについて，測ったりデータを取ったりする。そのため，測定されたデータが，本当に測りたいものなのかどうかが問題となる。例えば，喜びという感情の強さを測るために，尺度を作ってデータを取ったとする。その尺度の得点が，本当に喜びの強さを表していればよい。しかし，全く別の，例えば忍耐強さを反映したものに過ぎなければ，その質問紙は喜びの強さを測るものとして不適切である。このように，妥当性の低さは研究を無意味なものにするため，心理測定では古くから妥当性を重要な問題の一つとしてきた。

ただし，妥当性は尺度そのものの性質ではない。アメリカ教育研究学会（American Educational Research Association; AERA），アメリカ心理学会（American Psychological Association; APA），教育測定全国評議会（National Council on Measurement in Education; NCME）による「教育・心理検査のスタンダード」の2014年版（AERA, APA, & NCME, 2014, p.11）によれば，妥当性という言葉は尺度使用の解釈に対して用いるもので，「尺度の妥当性」という言い回しは正しくないとされている。前述の例でも，忍耐強さを反映した尺度の得点を，喜びの強さの得点と解釈することが不適切なのである。同じ尺度の得点を，忍耐強さの尺度の得点として解釈するならば，妥当性に問題はない。また，妥当性は有無の問題ではなく程度問題であることや，不変なものではなく新事実の発見や社会条件の変化などに伴って変化するものであること（Messick, 1989 池田訳 1992）もしばしば指摘される。

心理測定における妥当性の概念は細分化されており，また歴史的にも変化してきた。これらは村山（2012）が詳細に論じているが，ここでは古典的な妥当性の区分と，最近の妥当性の考え方について簡単に述べる。

古典的に，心理測定における妥当性は，内容妥当性，基準関連妥当性，構成概念妥当性の 3 種類に分けられてきた。内容妥当性とは，あることを測る尺度またはテストの内容が，結論を引き出そうとしているものをどれだけよく表現しているか，ということである。例えば，数学的能力を測るテストを作ることを考える。数学的能力は，計算，幾何，論理などから成り立つと考えられるため，テストにはこれらを測る項目が偏りなく含まれている必要がある。もしテストの項目が計算問題ばかりであった場合，あるいは全く別の，例えば語彙力を問う問題ばかりであった場合，そのようなテストの得点は，数学的能力を測るものとしての内容妥当性が低いことになる。

基準関連妥当性は，尺度が測ろうとしているものを測る他の変数（基準変数）とどれだけ強い関係があるか，ということである。基準関連妥当性は，個人の将来の基準変数の値をどの程度よく予

測できるか，という予測妥当性と，個人の現在の基準変数の値をどの程度よく推定できるか，という併存妥当性に分けられる。例えば，入学試験は，受験者の入学後の成績という基準変数の値をよく予測する必要がある。ここで求められているのは予測妥当性である。これに対し，期末試験は，受講者の現在の理解度という基準変数の値をよく推定できる必要がある。ここで求められているのは併存妥当性である。

構成概念妥当性は，尺度が測ろうとしている概念（構成概念）を説明する理論に照らして適切であるか，ということである。構成概念を一つしか想定しなければ，この定義は妥当性の定義と同じように見える。そこで，複数の構成概念を想定し，類似した構成概念を測る尺度どうしの値が類似したものになるか，という収束的妥当性や，異なる構成概念を測る尺度どうしの値が異なるものになるか，という弁別的妥当性を，構成概念妥当性のサブタイプとすることがある。

古典的に以上の3種類から成り立つとされてきた妥当性であるが，1980年代以降，Messick（1989 池田訳 1992） が提唱したように，妥当性は単一の概念であるという考え方が主流となる。すなわち，内容妥当性も基準関連妥当性も構成概念妥当性であり，構成概念妥当性は妥当性そのものである，という考え方である。その上で Messick（1995）は，構成概念妥当性を整理するため，次の6つの側面について説明している。すなわち，内容の側面，実体の側面，構造の側面，一般化可能性の側面，外的側面，結果の側面である（表3）。

内容の側面とは，尺度の内容が，尺度で測りたい領域と関係があるか，尺度で測りたい領域を代表するものであるか，ということである。

実体の側面とは，尺度に対する反応や回答のプロセスが，理論に合致しているかということである。

構造の側面とは，尺度得点の構造が理論的なものに合致しているかということである。

一般化可能性の側面とは，尺度の内容を，尺度で測りたいことに一般化できるかどうかということである。

外的側面とは，他の変数との収束的または弁別的な関係である。

結果の側面とは，測定結果の解釈が，どのような行動や影響，結果に結びつくかである。

以上の6つの側面は，妥当性を検証するためにどのような証拠が必要になるかの指針となる。「教育・心理検査のスタンダード」の2014年版（AERA, APA, & NCME, 2014, pp.13-21） では，妥当性の証拠として，内容に基づく証拠，反応プロセスに基づく証拠，内的構造に基づく証拠，他の変数との関係に基づく証拠，測定の結果と妥当性のための証拠，の5つを列挙している（妥当性の一般化は，他の変数との関係に基づく証拠に含めている）。これら全てについての証拠を集めれば妥当性が認められるのかといえば，そうではない。前述のように，妥当性は程度問題である。また，6つの側面は，妥当性という単一の概念を多面的に見るものである。したがって，これらを強迫的に満足させようとするのではなく，必要に応じて尺度の使用を批判的に検討するきっかけとするのが適切と考えられる。このように，妥当性を程度問題として捉えるべきであるとすることから，Messick（1995）による妥当性概念を単一のものとして捉える考えが普及し，古典的に妥当性概念を分割する考えから脱却しつつある。本論文でも，チェックリスト作成の際に妥当性概念を単一のものとして捉える立場から解説する。

村山（2012）は，心理測定における妥当性の問

表3　心理測定における妥当性の側面

側面		説明
内容の側面	Content Aspect	内容が，測りたい内容と関係があるか
実体の側面	Substantive Aspect	反応や回答のプロセスが，理論に合致しているか
構造の側面	Structural Aspect	尺度得点の構造が理論的なものに合致しているか
一般化可能性の側面	Generalizability Aspect	内容を，測りたいことに一般化できるか
外的側面	External Aspect	他の変数との収束的または弁別的な関係
結果の側面	Consequential Aspect	測定結果の解釈が，どのような行動や影響，結果に結びつくか

Messick（1995）をもとに作成。

題としてさらに，内容の幅の広い項目群による尺度作成，個人内相関と個人間相関の区別，尺度の不変性についても論じている。本稿では省略するが，これらも妥当性に関する重要な問題であるため，参照されたい。

4．チェックリスト

4.1　妥当性概念の整理

前章では，政策評価および，心理測定における妥当性概念について解説をした。政策評価においては評価指標の妥当性，心理測定においては質問紙における尺度構成の妥当性という違いはある。しかし，これらを比較すると，異なった文脈の下に妥当性の検討がなされており表現に違いはあるのだが，妥当性を構成する基本的要素自体は，共通している点が多い。それらの共通点を整理するために，表１に示した政策評価における評価指標が満たすべき条件に，表３に示した心理測定における妥当性の側面のいずれが関連しているかを検討する。以下，表３の心理測定における妥当性の６項目と表１の関係を，心理測定における妥当性の側面ごとに整理していく。

「内容の側面」は,「内容が測りたい内容と関係があるか」というものである。古典的区分の内容妥当性を，構成概念妥当性の一側面としたものといえるだろう。例えば，あるテストで数学的能力を測りたいとき，数学と無関係な語彙力を問う問題が含まれていれば，そのテストの得点は数学的能力を表すものとして不適切，つまり妥当性が低いといえる。また，掛け算についての問題として，掛けられる数が１になっている問題１問しかなければ，その解答は掛ける数と同じとなる特殊なケースと考えられ，掛け算を代表する問題と言い難く，妥当性を低めることになる。これは，指標が測るべき内容を測定しているかという点で，「1–1妥当性」の「計測すべき事象を計測」，と関連しているといえる。また，指標が目的に合致しているかという点，成果を適切に測定しているかという点で,「1–5目的との適合性」の「目的や成果を適切に反映」とも関連しているといえる。

「実体の側面」は,「反応や回答のプロセスが，理論に合致しているか」というものである。心理測定においては，反応や回答が実体の裏付けを持ったもので，偶然のものではないということを判断する側面といえる。例えば，数学的応用力を測りたいテストでは，計算のためにある程度時間がかかるものと考えられる。しかし，少なからぬ受検者が全ての問題にほぼ一瞬で答えていたとしたら，何らかのヒントがあった，当て推量でマークシートの特定の列だけにマークした，正答を不正に入手していた，などの理由が考えられ，その受検者の得点が数学的応用力を反映しているとは考えにくくなる。これも，指標が結果的に対象を測定するのに適したものであるということではなく，目的との結びつきに，論理的な説明が可能かという点で,「1–5目的との適合性」の「目的や成果を適切に反映」と関連しているといえる。

「構造の側面」は,「尺度得点の構造が理論的なものに合致しているか」というものである。例えば，ある尺度について，下位尺度が４つあり，それらは互いにある程度高い相関関係があると想定しているとする。それにも関わらず,1つだけ他の３つと全く無関係であるとしたら，合計得点は，測ろうとしているものと異なるものを表している可能性がある。これは，ある尺度の下位尺度が存在することが前提となっており，それらの内的一貫性，項目間での相関関係などが問題となる。これらの判断の手続きとして，因子分析を適用して因子構造を分析したりすることから，多数の質問項目が存在する質問紙設計において配慮すべき側面と捉えられる。それゆえ，今回扱っている指標設定の妥当性においては，該当する概念は存在しない。

「一般化可能性の側面」は,「内容を，測りたいことに一般化できるか」というものである。例えば，計算問題ばかりのテストでは，その得点の大小を，数学的能力の大小に一般化して解釈することに疑問が呈されるであろう。平井（2006）や村山（2012）は，従来は妥当性と並んで議論されてきた信頼性（同じものを測ったときに，同じ値が得られるかどうか，という性質）も，妥当性の一般化可能性の側面に含まれると指摘している。これは，Campbell and Stanley（1963）における外的妥当性の概念とも合致するものであり，同じ人に同じテストを再び行った場合に同じ値が得られるか，という再検査信頼性とも関係しているといえる。よって，この側面は「1–2信頼性」の「正確に計測」の概念の一部を表しているといえる。

「外的側面」は,「他の変数との収束的または弁別的な関係」というものである。古典的区分における構成概念妥当性で述べた，収束的妥当性や弁別的妥当性がこれに当たる。AERA et al.（2004），平井（2006），村山（2012）は，古典的区分の基準関連妥当性を，この外的側面に含めている。具体的には複数の質問項目が合った場合に，構成概念として近接した項目間では相関が高くなり，離れた項目間では相関が低くなると捉えられる。例として，国語の試験の妥当性を測る場合に，国語と英語の試験の成績は言語能力という点である程度の相関があることを想定することや，国語と数学は異なった能力を測定していることから，その相関は国語と英語の相関ほど高くないであろうことを想定することである。このように，外的な他の変数との関係という側面から妥当性を判断するものであり，直接的に対応する概念は表１に存在しない。この理由は，質問紙の項目設計では，複数回の調査を行い厳密に相関を見ながら精査して行くのに対し，評価指標の設計において，そのようなプロセスは現実的でないという，枠組みの違いによるものと思われる。その一方で，この考えを複数指標がある場合の判断基準と捉えると，他の変数との弁別的な関係の部分が,「2–2非重複性」の「重複無く異なる側面を計測」と関連しているともいえる。具体例を挙げると,「リーダーシップ力」を測る指標をいくつかの学生調査のアンケート項目をもとに作成してみたところ,「プレゼンテーション力」,「説得力」を示す項目と同じような項目が並び，その値もほぼ同じ様になっていることから妥当性が低い，と判断するような場合が考えられる。

「結果の側面」は,「測定結果の解釈が，どのような行動や影響，結果に結びつくか」というものである。これは測定内容そのものではなく，測定の結果が社会的に及ぼす影響を示している。例えば，生徒の学力を測定した結果，その値が生徒の学力向上に活用されればよい。しかし，測定に用いられたテストが特定の人種に不利で，その人種の生徒の学力を正しく反映できず，その生徒が適切な教育を受けられなくなってしまったとしたら，この測定結果の解釈は，結果の側面で妥当でないといえる。このように,「2–4反抗的行動への耐性」の「意図に反する行動を誘発せず」と合致していると捉えられる。また，指標の利用における意思決定者の政治的な配慮の必要性を示しているという点では,「1–10意思決定への有用性」の「意思決定に有益な知見を提供」とも関連しているといえる。

なお，表１においては，妥当性だけではなく信頼性も含まれており，心理測定における信頼性との対応関係も考えると，先に示した一般化可能性の側面に加え,「1–3理解可能性」の「意味が明確で誤解しにくい」は，心理測定における例としては「質問項目に誤解が無く，質問内容が正確に伝わるか」という，安定性の意味で，信頼性の概念の一部を表しているといえる。また,「1–9操作可能性」の「計測値の操作可能性が小さい」についても，心理測定においては，ノイズや人為的な操作により計測値がゆがめられず，同一の反応は同一の値が測定されるという安定性の問題と捉えられ，これも信頼性の一部といえる。

以上の様に，心理測定と，政策評価の妥当性，信頼性との関係について整理した。このように，妥当性はいくつかの基本的な共通要素があり，測る対象に応じて配慮すべき側面が異なると解釈できる。前章の心理測定の妥当性概念の捉え方を支持すると，妥当性は程度問題であり，いくつかの観測可能な側面から，妥当性という単一の概念を多面的にチェックして行く必要があると考えられる。以降では大学評価の文脈を重視し，そこでの妥当性を判断するためのチェックリストについて検討を行った内容を紹介する。

4.2 チェックリストの設計

ここでは，チェックリストの設計をどのように行ったかについて，詳細に解説する。チェックリストの基本となる考え方は，前節までの妥当性の基本的概念のレビューから，異なる専門分野であっても共通するものを取り出し，大学評価の文脈への適用可能性を検討した。また，現場担当の事務職員にも使い易い，平易な言葉での表現と，わかりやすい解説を加えたものであることを重視した。まず，表１に示した政策評価におけるチェックリストをベースに検討を始めた。その理由は，3.2.1節において説明したように，政策評価においては，現場の評価担当者向けの評価指標作成という視点の，テキストやマニュアル類にもと

づいて作成されたことから，それらの蓄積に基づいた平易な表現が主となっていたからであった。なお，表１では個別指標，指標群の区分がなされていたが，複数指標を想定するかどうかの議論を含めると複雑になり，一般向けには理解が難しくなることと，大学評価においては１つの目標・計画に複数指標が想定されることが一般的であるため，ここではそれらの区分を設けないこととした。

これをもとに，評価機関，研究機関，大学教員，事務職員という様々な立場のいずれからでも理解でき，指標作成の業務を想定して有益であると判断されるものに，チェックリストを絞り込んだ。また，説明についてもより平易で具体的なイメージが可能な内容へと変更を行った。

その中で，本来の指標の性質そのものの妥当性に関する点と，指標収集に関する現実的な側面に関する点に整理可能であったため，それらの区分を設けることとした。さらに，いくつか重複するとも捉えられる内容について，まとめて１つの項目にするなどの整理を行った。また，「1–6施策の影響の大きさ」は「計測事業への施策の影響度」という内容から，目的と指標の内容ではなくその前の段階の事業と施策の関係に関連するチェックリストであるため，今回の検討からは除外した。

以上の作業プロセスを示した結果，導かれた項目および，表１と表３との関係を表４に示す。妥当性として５項目，実用面として５項目のチェックリストを作成した。以下，基準ごとに解説をして行く。

妥当性に関する最初の基準である「目的との適合性」は「指標が，計画の進捗や目指す成果を適切に反映しているか」という内容で，妥当性において最も基本的かつ重要な内容といえる。これは，表１における「1–1妥当性」，「1–5目的との適合性」および，表３の「内容の側面」，「実体の側面」と関連している。大学評価において，目標・計画の成果や進捗を測定する指標を検討する際にも，当然ながら最も重視されるべき内容である。たとえば「ボランティア活動を推進する」という目的があった場合に「緊急災害時の支援体制（人数）」という指標があった場合，全く関連がないとは言えないが，緊急時の特殊な例という点で目的との適合性は低いといえるだろう。

「調査対象・結果への影響」は，「指標設定の結果，意図しない悪影響を及ぼすものではないか」という内容で，指標を設けたことにより，マイナスの影響が生じないように配慮すべきことを示した基準である。これは，表１の「2–4反抗的行動への耐性」，表３の「結果の側面」と関連している。後述する図１の教材の例にも示すように，大学評価においても，指標を設定することで，その向上のみが目的化し，結果的に負の影響が生じる可能性があることに，配慮した基準である。

「信頼性」は「誰がいつ測定しても，同じ事象

表４　本論文で提案するチェックリストと政策評価，心理測定における妥当性との対応関係

本論文で提案するチェックリスト		政策評価（表１より）	心理測定（表３より）
区分	基準	基準	基準
妥当性	目的との適合性	1–1妥当性 1–5目的との適合性	・内容の側面 ・実体の側面
	調査対象・結果への影響	2–4反抗的行動への耐性	・結果の側面
	信頼性	1–2信頼性	・信頼性 ・一般化可能性の側面
	理解可能性	1–3理解可能性	・信頼性
	包括性・非重複性	2–1包括性	
		2–2非重複性	・外的側面
実用面	意思決定者への有用性	1–10意思決定への有用性	・結果の側面
	計測可能性	1–7計測可能性	
	収集の適時性	1–4タイムリー性	
	データ収集のコスト	1–8データの収集費用 2–3データの収集費用	
	操作可能性	1–9操作可能性	・信頼性

や状態からは同じ測定結果が得られるか」という内容である。本論文で提案するチェックリストでは，3.2.2. 節の一般化可能性のように，広い意味では信頼性も妥当性に含める立場もあること，および区分を増やすことで複雑さが増えることを避けるために，信頼性も妥当性チェックリストの中に置くこととした。これは，表１の「1–2信頼性」，表３の「信頼性」，「一般化可能性の側面」と関連している。大学評価の文脈における具体例としては，本部の人間が，部局や担当部署に同一のデータを再度要求しても，収集時にばらつきが生じたり，大きく異なるようなことのない測定法で得られているか，という内容に相当する。

「理解可能性」は「指標の意味が，明確でわかりやすく，誤解が生じないか」という内容で，指標のわかりやすさ，理解の正確さと関連するものである。これは，表１の「1–3理解可能性」，表３の「信頼性」と関連しているように，信頼性の一部も含まれていると考えられる。大学評価の文脈における具体例としては，「論文数を指標として研究費配分を行う」という場合に，一概に論文と言っても，A 学部では査読付き海外雑誌のみ，B 学部では査読無し紀要も含む，C 学部では国際会議プロシーディングスを含む等，様々な定義がある場合があり，誤解のない様に細かく定義を定める必要性があることをチェックする基準である。

「包括性・非重複性」はそれぞれ分離することも可能な基準である。「包括性」は「計画の重要な側面が，もらさず指標によってカバーされているか」であり，「非重複性」は「指標間に重複がなく，各指標は異なる側面を計測しているか」という内容である。これらは複数指標が存在する場合の指標間の関係性について述べている点を考慮し，出来るだけ基準の数を増やさないほうが良いという考えに基づき，組み合わせて１つの基準とした。これは，それぞれ表１の「2–1包括性」，「2–2非重複性」と対応しており，非重複性は，前節に示した意味で，表３の「外的側面」と関連している。大学評価においても，「包括性」は，指標に漏れが無いか，「目的との適合性」とも関連して必要充分なものが収集されているかということをチェックする基準といえる。また，「非重複性」は，省力化・効率化の観点から同じような指標を複数設定することのないよう，他の指標と区別が可能かを考慮して絞り込むべきことをチェックする基準である。特に，本部において評価を遂行する上では，部局の負担軽減への配慮と信頼関係の確立を考慮すると，非重複性は重要といえる。

次に，実用面の基準について解説する。「意思決定者への有用性」は「指標が，執行部等の意思決定者に対して，有益な知見を提供してくれているか」という内容である。表３の「結果の側面」は，「調査対象・結果への影響」とはネガティブな側面での意図しない結果を誘発しないかを防止するというものであったが，逆にポジティブな側面として結果を捉えると，この基準と関連しているといえる。大学評価においては，一般に評価室，IR 室等を考えると担当理事，センター長等の意思決定に利用可能な指標でないと，理論的に導かれたとしても実用面では意味を持たないという点で，重要な基準といえる。

「計測可能性」は，「指標となるデータは収集可能か」というものであり，これは表１の「1–7計測可能性」と同義である。これは現実面の話であり，表３に対応する概念は無い。大学評価においても，理想的には計測したい指標として想定されるものでも，倫理的な側面や現実的に収集不能なものである指標だと，実用的には問題があるということをチェックする基準である。

「収集の適時性」は，「有用なタイミングで，指標の計測値は入手可能か」というものであり，これは表１の「1–4タイムリー性」と同義である。これも，計測可能性と同様，表３に対応する概念は無い。用語について修正を施したのみである。大学においては，法人評価や認証評価等の外部の評価時のエビデンスとしてや，補助金申請時の提出データ等，それぞれの必要な時期に入手可能なものであるかが重要であることから，本基準を設けた。

「データ収集のコスト」は，「データを収集するための費用は大きすぎないか」というものであり，これは表１の「1–8データの収集費用」「2–3データの収集費用」と同義である。これも表３との対応関係は，上の２つの基準同様，存在しない。大学における指標設定時に，既存のデータからの設定が難しい場合，新たにデータ収集を考える必要性がある場合が想定される。そのような際に，実用的な側面からは，コストについて考慮することは必須であろう。例えば，卒業生，在校生，企業等

へのアンケート調査等によるプログラムの効果や学習成果の調査は良く行われているが，大学側や調査対象者の負担を考慮すると，精査して行う必要があるといえるだろう。

「操作可能性」は，「指標の計測値は，都合良く操作して変更可能なものではないか」というものであり，これは表１の「1–9操作可能性」と同義であり，表３の「信頼性」と関連している。大学で考えると，例えば，全学的に各部局に調査を行った場合，解釈により都合の良いような数値へと操作することで有利な結論を導きだすことが可能な状況であると，正確な把握ができないという問題が生じることとなり，公平性や正確さの確保という点から配慮すべき基準といえる。

4.3　チェックリストとその解説

前節に示したようなプロセスを経て，最終的に整理したチェックリストとその解説を表５に示す。

なお，文章のみの説明では，一般向けには理解が難しいと思われるものがあることから，具体的な事例や直感的にイメージ可能なように図を挿入した教材を開発した。一部を抜粋して示す（図１）。全体については，大学評価・学位授与機構の Web ページ（シンポジウム・セミナー）から参照可能である。

図１に示した例は，「調査対象・結果への影響」の基準を説明するスライドである。説明の教材は，タイトルに基準を示し，いくつかのチェックした際に問題となる事例を示すものであった。ここでは，4.1節にて示した結果の側面における牛乳パックを用いたリサイクルの例および，学生の学力向上のための現状把握として学力テストの平均得点を指標とすることが，結果としてテストの得点自体が目的となり，一部の学力の低い学生を欠席させる様になったという例を挙げている。

妥当性，実用面の区分および各基準とその説明を示す。

「調査対象・結果への影響」のチェックリスト基準を例として，教材のスライドを示す。内容の解説および二つの例と，注意すべき内容について示している。

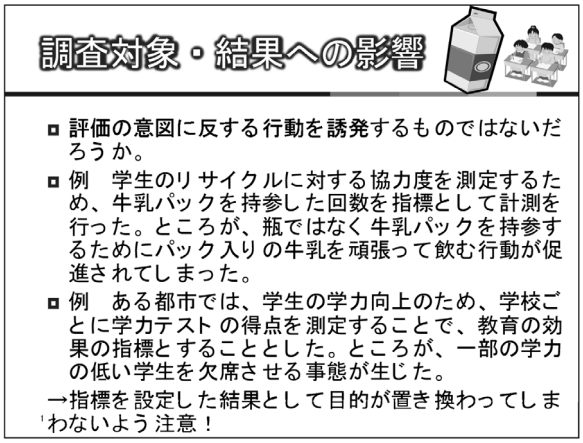

図１　教材の例（調査対象・結果への影響に関する妥当性の解説）

表５　完成したチェックリストと解説

区分	基準	説明
妥当性	目的との適合性	指標が，計画の進捗や目指す成果を適切に反映しているか。
	調査対象・結果への影響	指標設定の結果，意図しない悪影響を及ぼすものではないか。
	信頼性	誰がいつ測定しても，同じ事象や状態からは同じ測定結果が得られるか。
	理解可能性	指標の意味が，明確でわかりやすく，誤解が生じないか。
	包括性・非重複性	計画の重要な側面が，もらさず指標によってカバーされているか。
		指標間に重複がなく，各指標は異なる側面を計測しているか。
実用面	意思決定者への有用性	指標が，執行部等の意思決定者に対して，有益な知見を提供してくれているか。
	計測可能性	指標となるデータは収集可能か。
	収集の適時性	有用なタイミングで，指標の計測値は入手可能か。
	データ収集のコスト	データを収集するための費用は大きすぎないか。
	操作可能性	指標の計測値は，都合良く操作して変更可能なものではないか。

5．おわりに

本稿では，一般的・包括的な指標設定に関する妥当性を判断するためのチェックリストを作成し，大学での利用について検討した。学術的には妥当性についての多くの研究があるが，複数分野のレビューを通して比較検討することで，本質的な点では多くの部分で共通していることがわかった。

大学内で評価の実質化を進めるためには，評価に携わる教員，事務職員，そして評価関連部署だけではなく教学の現場に携わる事務職員など，幅広く異なる立場における人々の協力が不可欠である。特に，日頃の業務において，エビデンスとしての指標収集の意識の向上および，その妥当性についての理解を広めて行く必要があると考えられる。そのため，いずれの立場においても，理解し易い内容を持ったチェックリストとその解説が望まれる。本論文で作成したチェックリストは，学術的な妥当性概念の一般的かつ包括的な整理を行い，指標妥当性の大学評価の場への利用を検討したはじめての取組といえる。

今後の課題として，今回作成したチェックリストは，妥当性を判断する１つの目安となってはいる一方で，概念的な整理はされているが，解釈の幅が広い問題がある。そのため，利用する側から見て，評価の対象に応じて専門性を持った解釈が必要となり，誰もが機械的に判断できるというものではない。これらの点を解決するために，例えば，学内組織の教務部門向け，総務部門向けといった業務内容による細分化や，全学組織向け，部局担当者向けのような組織階層による細分化など，チェックリストの説明とその事例をブレイクダウンして行く必要がある。

また，妥当性の各基準の学術的に厳密な説明と，使いやすさや理解のしやすさとの間にトレードオフが生じることも，今回のチェックリスト作成の際に浮き彫りとなった問題である。有意義で使い易いチェックリストへ向け，その程度の調節をして行く必要がある。

試行的な実践として，平成27年１月29日に「指標の選び方＆指標信頼性・妥当性のチェックリスト」というワークショップを大学関係者36名に対して行い，その中で本ツールについての紹介を行った（大学評価・学位授与機構，2015）。それらのアンケートの集計結果を見ると，高い満足度が見て取れる。ただし，それがチェックリストの本来の目的である，大学現場の業務支援にどの程度寄与するのかは，実際の活用状況をトレースしてみてゆく必要もあるだろう。

今後はツールとしての利用の普及を充実するために，更なる実践と，具体的な事例の充実や参加者アンケートを通した改善を続け，幅広く有用なものとして提供して行く予定である。以上，本稿が大学の自己評価能力向上の支援へ向けた一助となると幸いである。

引用文献

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

American Society for Public Administration（2000）. *Performance Measurement: Concepts and techniques.* Washington. D.C.: ASPA Center for Accountability and Performance.

Ammons, D.N.（1996）. *Accountability for Performance: Measurement and Monitoring in Local Government.* Washington. D.C.: Internation City/Count Management Association.

Campbell, D.T. & Stanley. J.（1963）. *Experimental and quasi-experimental designs for research on teaching.* In N. L. Gage（Ed.）, Handbook of research on teaching（pp. 171-246）. Chicago, IL: Rand McNally.

Chen, H. T., Donaldson, S. I., & Mark, M. M.（2011）. *Validity frameworks for outcome evaluation.* In H. T. Chen, S. I. Donaldson, & M. M. Mark（Eds.）, Advancing validity in outcome evaluation: Theory and practice. *New Directions for Evaluation*, 130, 5-16.

大学評価コンソーシアム（2013）．データ収集作業のガイドライン―効率的・効果的な評価作業のためのデータ収集の課題と対応―.

第３期中期目標期間における国立大学法人運営費交付金の在り方に関する検討会（2015）．第３期中期目標期間における国立大学法人運営費

交付金の在り方について審議まとめ .
独立行政法人　大学評価・学位授与機構　シンポジウム・セミナー　EA ワークショップ「指標の選び方＆指標信頼性・妥当性のチェックリスト」< http://www.niad.ac.jp/n_kenkyukai/1259551_1207.html >（2015年 6 月30日アクセス）
独立行政法人　大学評価・学位授与機構（2013a）. 大学の内部質保証力を向上させるための支援ツールの開発と普及に関する報告書 .
独立行政法人　大学評価・学位授与機構（2013b）. 実績報告書作成要領　国立大学法人及び大学共同利用機関法人の第 2 期中期目標期間の教育研究の状況についての評価.
独立行政法人　大学評価・学位授与機構（2014）. 大学期間別認証評価　自己評価実施要項（平成27年度実施分）.
Guba, E. G., & Lincoln, Y. S.（1994）. *Competing paradigms in qualitative research.* In N. K. Denzin, & Y. S. Lincoln（Eds.）, Handbook of qualitative research（pp. 105-117）. Thousand Oaks, CA: Sage.
Hatry, H.P.（1999）. *Performance Measurement: Getting results.* Washington, DC: Urban Institute Press.
Hatry, H.P.（2006）. *Performance Measurement: Getting results*（2nd ed.）. Washington, DC: Urban Institute Press.
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（編著）心理学研究法の新しいかたち，誠信書房，pp.21-49.
JICA 評価部（2013）. 無償資金協力開発課題別の標準指標例 Ver.2.
北川剛司（2008）. 教育評価における妥当性・信頼性に関する一考察，広島大学大学院教育学研究科紀要，57，99-104.
Langbein, L.（2012）. *Public Program Evaluation: A Statistical Guide* 2nd Edition. New York: M.E.Sharpe.
Messick, S.（1989）. Validity. Linn. In R.L.（Ed.）*Educational Measurement,* 3rd ed. New York: McMillian, pp.13-103.
（メシック，S. 池田 央（訳）（1992）. 妥当性. リン，R.L.（編）池田 央・藤田恵璽・柳井晴夫・繁桝算男（監訳）教育測定学 原著第 3 版上巻，C.S.L. 学習評価研究所，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（9）, 741-749.
文部科学省（2015）. 平成27年度大学教育再生戦略推進費「地（知）の拠点大学による地方創生推進事業（COC ＋）」公募要領 .
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，51, 118-130.
成田滋（1986）. 単一被験者実験計画法における内的妥当性と外的妥当性に関する一考察，特殊教育学研究，23（4），37-44.
大津起夫（2011）. 信頼性と妥当性. 松原 望（編）統計応用の百科事典，丸善出版，pp.426-427.
大塚雄作（2007）. 大学教育評価における評価情報の信頼性と妥当性の検討，工学教育，4, 14-20.
渋井進・面高俊宏（2009）. 国立大学法人評価の実績報告書の作成プロセス―地方総合大学における事例―，大学評価・学位研究，10，47-58.
田中啓（2014）. 自治体評価の戦略 ―有効に機能させるための16の原則，東洋経済新報社.
田中弥生（2009）. 評価可能性のアセスメント（Evaluability Assessment）～大学の自己評価能力向上のために～，大学評価・学位研究，10，27-44.

（受稿日　平成28年 1 月15日）
（受理日　平成28年11月25日）

[ABSTRACT]

Checklists for setting indicators to enhance the self-assessment capability of universities

SHIBUI Susumu*, ASANO Shigeru**, HASHIMOTO Takamitsu***, ONO Hiromu****
ONO Tatsuya*****, YAMASAKI Sono******, TANAKA Yayoi*******

The National Institution for Academic Degrees and University Evaluation (NIAD-UE) established the "Evaluability Assessment (EA) Study Group" to attempt to improve university evaluation practices in Japanese universities. In this article, we present a checklist that was developed to set better performance indicators in university evaluation. To meet the requirements of policymakers in the distribution of public funding, Japanese universities should demonstrate their performance with quantitative data. This circumstance has increased the importance of developing performance indicators, and the University Evaluation Office or Institutional Research (IR) Office is expected to create them. This office also needs to ensure the validity of the indicators; however, this is not an easy task. To ease this problem, our study group created a simple checklist. We will start with examples that show the problem with the current university evaluation and how indicators are developed and used. Then, we will discuss the checklist while reviewing the previous studies on administrative evaluation and psychological measurement.

* Associate Professor, Research Department, National Institution for Academic Degrees and Quality Enhancement of Higher Education
** Professor, Academic Assembly, Yamagata University
*** Assistant Professor, Research Division, National Center for University Entrance Examinations
**** Associate Chief Administrative Officer, Strategic Planning Division, Kwansei Gakuin University
***** Professor, Faculty of Regional Seiences, Tottori University
****** Deputy Director, Kyoto University of Foreign Studies, Division for the Coordination of Planning and Implementation
******* Professor, Research Department, National Institution for Academic Degrees and Quality Enhancement of Higher Education