第5分科会

高次の能力を捉えるための評価
～どのような評価がどのような能力を捉えることに適しているのかを課題づくりも含めて考える～

報告者

大塚　雄作　　独立行政法人大学入試センター　試験・研究統括官／教授

松下　佳代　　京都大学　高等教育研究開発推進センター　教授

コーディネーター兼報告者

斎藤　有吾　　京都大学　高等教育研究開発推進センター　特定助教

【参加者　67 名】

昨今、アクティブ・ラーニングの趨勢とともに、高次の（統合的な）能力の評価法に関する議論が多く行われている。しかし、例えばルーブリックなど評価基準への注目は集まっているが、当該の能力を可視化するための課題の作成やその良し悪しの検討といったことに関する議論はまだ発展の途上にある。

そこで、日本の高等教育研究における教育評価の第一人者とともに、高次の能力を捉えるための評価と課題づくりに関して議論する。

第5分科会

高次の能力を捉えるための評価
～どのような評価がどのような能力を捉えることに適しているのかを課題づくりも含めて考える～

京都大学　高等教育研究開発推進センター　特定助教　斎藤　有吾

概要

昨今、アクティブ・ラーニングの趨勢とともに、高次の（統合的な）能力の評価法に関する議論が多く行われている。しかし、例えばルーブリックなど評価基準への注目は集まっているが、当該の能力を可視化するための課題の作成やその良し悪しの検討といったことに関する議論はまだ発展の途上にある。そこで、本分科会では、日本の高等教育研究における教育評価の第一人者である、大塚 雄作 独立行政法人大学入試センター試験・研究統括官／教授と松下 佳代 京都大学高等教育研究開発推進センター教授とともに、高次の能力を捉えるための評価と課題づくりに関して議論した。

1.分科会のねらいと問題意識

昨今の高等教育では、学問の基本的な知識を獲得するだけでなく、知識の活用能力や創造性、生涯を通じて学び続ける能力などを培うことが重視されている。また、それだけでなく、それらを学習成果として客観的な指標で捉えて可視化することが求められている。近年では、高大接続の視点で入試改革の議論にも発展している。詳細は学士課程答申、質的転換答申、高大接続答申（中央教育審議会, 2008；2012；2014）を参照されたい。

例えば学士力の「統合的な学習経験と創造的思考力」は、「これまでに獲得した知識・技能・態度等を総合的に活用し、自らが立てた新たな課題にそれらを適用し、その課題を解決する能力」（中央教育審議会, 2012, p.13）とある。これは知識・理解のレベルを超えた高次の（統合的な）能力であり、このような能力をどのように捉えることができるのかという議論を進めていく必要がある。昨今の高等教育における学習成果の可視化の方法を概観すると、GPA、教育関連企業が開発した汎用的技能や社会人基礎力を評価するためのテスト、学生調査などが多用されている。しかし、それらの方法で、大学のディプロマ・ポリシーに対応するような、知識や理解のレベルを超えた能力を捉えることはできるのだろうか。なお、これらさまざまな学習評価を整理するものとして松下（2017）の＜直接評価・間接評価＞、＜量的評価・質的評価＞、＜科目レベル・プログラムレベル・機関レベル＞といった３つの軸による分類が参考になる。合わせて参照されたい。

さて、これまで多用されている評価以外の方法で、特に高次の（統合的な）能力を捉えようとする動きもいくつか挙げることができる。例えば、科目レベルのパフォーマンス評価をプログラムレベルの学習成果の可視化のための指標として用いようとする新潟大学歯学部の事例（小野, 2017）や、大学教員がテスト問題を共同で作成するとともに、作成したテスト問題を広く共有することを通して、コンピテンスと学習成果についての対話を喚起し、共通理解を形成することを目指す取り組み（Tuningテスト問題バンク；深堀, 2017）などが挙げられる。この２つはどちらも高次の（統合的な）能力を捉えようとする試みだが、前者は標準化を志向せず、後者は標準化を志向しているという違いがある。本分科会企画者兼コーディネーターである斎藤 有吾 京都大学高等教育研究開発推進センター特定助教は、その２つの取り組みに関わっているが、両者に共通して実感した問題点として、捉えようとする能力に対応する「評価課題」と「評価基準」を作成することは、当該の専門分野の大学教員でも困難であるということである。これらは本来ワンセットで議論されるべきことであるが、ルーブリックなど評価基準への議論や研究は多くなされてきている一方で、評価課題の作成に関する議論はまだ途上である。

そこで本分科会では、大規模標準テストなどに代表される量的評価の第一人者である大塚教授と、パフォーマンス評価などに代表される質的評価の第一人者である松下教授から、それぞれの立場から高次の（統合的な）能力を捉えるための評価、特に評価課題づくりに関して、ご講演を頂いた。また、講演の後に、２つの講演の内容を深めると共に、そのような評価課題の良し悪しを検討する視点を、異なる立場を超えて共有することを目的として、参加者が量的評価と質的評価の立場に分かれ、実際に高次の（統合的な）能力を捉えるための評価課題を作成するグループワー

クを行った。

2.分科会の構成と内容

第５分科会のタイムテーブルを表１に示す。まず、松下教授から、高次の（統合的な）能力とはどのような能力を指すのか、評価をする際にどのようなことを考えるべきなのか、学習評価における＜直接評価・間接評価＞、＜量的評価・質的評価＞という軸による分類、パフォーマンス評価やルーブリックに関する基本事項の説明が行われた。パフォーマンス評価は質的評価であり直接評価である。その後、パフォーマンス評価の３つの事例が紹介された。３つの事例はそれぞれ、パフォーマンス評価をデザインする際にポイントとなる①評価したい能力、②求めるパフォーマンス、③設定する評価課題、④設定する評価基準、⑤評価する主体などが、それぞれ異なるものであり、比較して検討することが可能なものであった。それらの中でも、デジテル・リテラシーの評価の事例は、知識の活用を把握することができ、しかも評価負担が小さい評価の開発に関することであり、講演後のグループワークに直結する情報が提供された。

次に大塚教授から、量的評価の視点から、良問とはどのような要件を満たすものなのかということや、記述式課題とその採点の問題点について、実例を交えて詳しい説明がなされた。例えば、大規模な標準テストは、量的評価が基盤とする教育測定学（テスト理論）の専門的知見からの検証が求められるものであり、それぞれの問題（項目）の特性や、試験全体から得られる尺度得点の信頼性・妥当性の検討などが欠かせない。また、「良問」は問題それ自体が決めるものではなく、試験全体の枠組みにおいて議論されるべき相対的なものである。このように、質的評価とは明らかに異なった視点から、特定の能力を捉えるための評価のポイントが明らかにされた。加えて、昨今の大学入試改革の議論において、知識・理解より高次の能力を捉えることを目的として導入が叫ばれている記述式課題に関して、採点（評価）者間の信頼性の担保の困難さや、採点にかかる時間やコストなどの問題点が指摘された。

このように参加者には、＜量的評価・質的評価＞の双方の立場から、評価課題や評価基準を作成する際に留意するべきこと

表１　第５分科会 タイムテーブル

	タイム枠	演目・登壇者等	内容・テーマ等
	10:00～10:10（10分）	開会挨拶・趣旨説明	第5分科会の趣旨説明
午前	10:10～11:00（50分）	第1発表：松下 佳代 先生	テーマ：高次の能力を捉えるための評価－パフォーマンス評価のデザイン－
	11:00～11:50（50分）	第2発表：大塚 雄作 先生	テーマ：テストをテストする視点～共通テストの良問とは～
	11:50～12:00（10分）	午後の趣旨説明と質問の記入	
	12:00～13:30（90分）	昼休み	
午後	13:30～13:40（10分）	質疑への応答	
	13:40～13:50（10分）	第3発表：斎藤 有吾	テーマ：高次の能力を捉えるための評価のデザインとグループワークの説明
	13:50～14:30（40分）	グループワーク	
	14:30～14:40（10分）	発表準備	
	14:40～15:10（30分）	発表と投票	
	15:10～15:25（15分）	3グループ発表と講評	
	15:25～15:30（5分）	ワークシート記入	

や、それらの良し悪しをどのように検討したらよいのかという視点が提供された。

そして午後の部のグループワークでは、参加者にデジタル・リテラシーを捉えるための評価課題を作成する課題が与えられた。量的評価の視点から、選抜試験に用いるという設定のもと、デジタル・リテラシーを捉えるパフォーマンス評価型の課題づくりを行う７グループと、質的評価の視点から、科目の終了時の総括的評価に用いるという設定のもと、パフォーマンス課題づくりを行う７グループに分けられた。どちらもグループメンバーは５名以内となるように調整された。

グループワークによって作成された課題（成果物）をポスター発表形式で発表してもらい、講演者２名の審査と参加者同士の投票によって、優れた課題を作成したグループには「大塚賞」「松下賞」「オーディエンス賞」が与えられた。各グループは、定められた条件設定のもとで、高次の（統合的な）能力を捉えるための評価をつくることの困難さを実感したようであった。また、講演者２名の講評によって、新たな気づきを得た参加者も散見された。加えて、一方の条件設定のもとで作成されて「良い」と考えられた課題が、もう一方の条件設定のもとでは必ずしも「良い」といえなくなる可能性など、＜量的評価・質的評価＞それぞれの視点を往還する必要性を意識する重要な契機となったようであった。

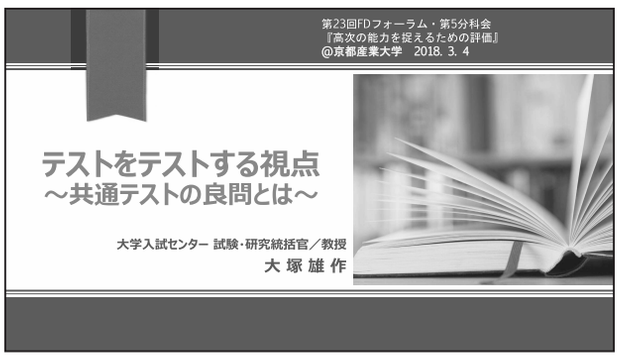
２名の講演者には、講演内容やグループワークに関して提案や議論をいただきつつ、分科会当日まで熱心に準備を進めて頂いた。また参加者の皆様には、拙いコーディネートにも関わらず、積極的にグループワークに関与していただき、その成果物からさらなる論点を提供していただいた。最後に、会場のスタッフの皆様の臨機応変な対応によって、トラブルなく分科会を終了することができた。紙上を借りて厚く御礼を申し上げる。

第５分科会コーディネーター
斎藤　有吾（京都大学　高等教育研究開発推進センター　特定助教）
（ugo.saito@gmail.com）

参考文献
中央教育審議会（2008）『学士課程教育の構築に向けて』（答申）
中央教育審議会（2012）『新たな未来を築くための大学教育の質的転換に向けて～生涯学び続け、主体的に考える力を育成する大学へ～』（答申）
中央教育審議会（2014）『新しい時代にふさわしい高大接続の実現に向けた高等学校教育，大学教育，大学入学者選抜の一体的改革について』（答申）
深堀聰子（2017）「エンジニアリング教育の達成度評価～テスト問題バンクの取り組み～第１回:連載にあたって：テスト問題バンクの活動について」『日本機械学会誌』120（1178）, 38-39.
松下佳代（2017）「学習成果とその可視化（特集 高等教育研究のニューフロンティア）」『高等教育研究』20,93-112.
小野和宏（2017）『パフォーマンス評価と教育の質保証－新潟大学歯学部の取組－』,東邦大学医学部FD講演会講演資料,2017年８月３日,アワーズイン阪急,東京.

テストをテストする視点
～共通テストの良問とは～

独立行政法人大学入試センター　試験・研究統括官/教授　大塚　雄作



1．良問の要件を探る

（1）テストスタンダードが定める要件
（2）試験の枠組みによる制約
（3）測定目的との関連性 ― 何を測りたいのか
（4）教育へのインパクト

2

（1）テストスタンダードが定める要件

1.1 テストの基本設計

開発者は**利用目的**や場面にあわせて，測定内容，測定形式，実施方法・手続，結果の利用方法，適用を想定する対象者の範囲などを明確に定め，**基本設計**を行う。

1.2 測定内容の定義と構造化

開発者は，測定しようとする能力，学力，性格，行動などの特性を明確に定義し，それが表現できるような**適切な尺度を構成**する。なお，測定しようとする特性が複数の下位の特性で構成される場合は，その構造を明らかにし，それらも測定できるように複数の下位尺度を設計する。

3

（1）テストスタンダードが定める要件

1.5 採点手続の設計

テストの開発においては、その採点手続を詳細かつ具体的に示すべきである。客観式テストにおいては、採点手続の理論的根拠を明らかにする。主観的評定においては、評定基準および評定手順を明確に設定する。

1.6 尺度化の方法

採点結果から尺度得点を求める場合は、さまざまな尺度化の方法の中から、最も適切な方法を吟味して選択する。その選択根拠は、求めに応じて説明できるように記録しておくべきである。

1.7 尺度の標準化

汎用されるテストは、規準とする集団を明確に定め、その集団における相対的位置づけによって尺度化することが望ましい（この手続を「標準化」という）。標準化においては、用いた標本と標準化手続について記録し、公開する、また、標準化の結果は定期的にその有効性を確認し、改訂の必要性の有無を検討する。

4

（1）テストスタンダードが定める要件

1.10 尺度得点の確からしさの推定と公開

開発者は、構成された尺度得点がどの程度安定しているかを、しかるべき統計指標を算出して検討し（この過程を「**信頼性**の確認」という）、その結果を公開すべきである。なお、テストが複数の下位テストから構成される場合は、それぞれの下位テストごとに検討し、その結果を公開すべきである。

1.11 尺度得点の適切さの確認

開発者は、構成された尺度が測定内容として定義された特性をどの程度適切に測定しているかを多面的に検討し（この過程を「**妥当性**の確認」という）、その結果を公開すべきである。

5

（1）テストスタンダードが定める要件

★特に大規模な標準的テストは　教育測定学等の専門的知見からの検証が求められる。

- **「信頼性」・「測定誤差」**
- **「妥当性」・「構成概念」**
- **「識別力」・「情報量」**
- **「尺度得点」・「標準化」**

etc.

★形成的評価においても、それに利用される評価情報、そのベースとなる測定値などについては、信頼性・妥当性・識別力などは、的確な評価に結び付くためには必要不可欠な要件と言える。

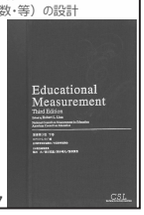
6

（2）試験の枠組みによる制約

■ 学力・能力テストの設計と開発

1. **テストの目的** ： テスト開発の諸側面 ＝ テストの目的に依存
2. **テストの設計** ： 外的背景要因（受験者集団・時間的制約・等）を踏まえ、内的テスト属性（項目内容・項目の型・採点基準・項目数・等）の設計
3. **項目の開発** ： 項目の型を定め、具体的に項目を作成
4. **項目の評価** ： 難易度・識別力等の統計的項目評価
専門家による項目の内容的評価
5. **項目の選択** ： 一部の統計指標のみならず、識別力・迷わし分析、専門家の内容評価など、総合的観点から選択
6. **テストの構成** ： 項目の配列・項目のレイアウト・等

＊J.ミルマン・J.グリーン（1989）．大塚雄作（訳）（1992）．第8章 学力・能力テストの設計と開発 R.L.リン（編）『教育測定学・下』, 3-53．C.S.L.学習評価研究所 より



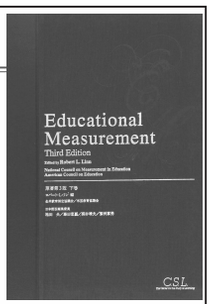
7

（2）試験の枠組みによる制約

■ テスト開発計画におけるいくつかの問題点

◆**外的な文脈的要因**
テストの**目的**は何か？
だれがテストを受けるか？
どの程度の**テスト時間**が利用可能か？
どのようにテストが施行されるか？

◆**内的なテスト属性**
テストが対象とする範囲は何か？
項目の形式としてどのようなタイプのものが利用されるか？
どれだけの数の項目を作成するか？
どのような**項目**の計量的特徴が望まれるのか？
項目はどのように評価され選別されるか？
項目や**テスト**はどのように得点化されるか？
どのような**テスト**の計量的特性が望まれるのか？



＊J.ミルマン・J.グリーン（1989）．大塚雄作（訳）（1992）．第8章 学力・能力テストの設計と開発 R.L.リン（編）『教育測定学・下』, 3-53．C.S.L.学習評価研究所 より

8

（2）試験の枠組みによる制約

★ テストの枠組みをどう定めるか、 外的な文脈的要因による制約は何かによって、 テストに含まれる 良問の評価は 違ってくる。

★ ある問を良問とするためには、 テストの枠組みや 外的な文脈的要因を それに合わせて変更する必要がある。

★ 適切な枠組みを設定することによって、 目的に応じた「測定＝テスト」が保障されることになる。

9

（3）測定目的との関連性 ― 何を測りたいのか

●学力・能力テスト（調査）の目的のいろいろ

テストの対象と目的		教育課程（カリキュラム）			対 応
		前	中	後	
個人	診断・指導	配置 診断的評価	診断 形成的評価	成績 総括的評価	個人的指導・保護者等への報告
	選抜・資格	選抜 入学試験	ガイダンス	進級 資格・検定試験	認定・資格付与
集団	調査・研究	教育課程前 の状況把握	教育課程 の評価 全国学力調査	教育課程 の効果把握 PISA調査	全体的状況報告（国・県・学校等）

＊J.ミルマン・J.グリーン（1989）．大塚雄作（訳）（1992）．第8章 学力・能力テストの設計と開発 R.L.リン（編）『教育測定学・下』, 3-53．C.S.L.学習評価研究所 より改編

10

（3）測定目的との関連性―何を測りたいのか

★ 選抜試験と調査の違い を認識すべきである
- ➢選抜試験：個人差を識別する → 個人の尺度得点の信頼性
→ 十分な項目数の準備等が必要
- ➢調　　査：集団の特徴を知る → 集団における得点の信頼性
→ 偏りのない十分な標本サイズの確保

★ 総括的評価と形成的評価の違い を認識すべきである
- ➢総括的評価：尺度得点等による数量的指標
（集団における相対的位置を表す標準化指標など）
- ➢形成的評価：個人個人の学習状況・個性に応じた質的表現
（パフォーマンスを教師等が把握して改善のためにアドバイス）

11

（4）教育へのインパクト

▪ **評価の流れ**
Input → Process → Output → Outcome → **Impact = washback effect**

▪ **テストのインパクト ＝ 波及効果 washback effect ･･･ 波及のあり方は多様**
- ➢**「妥当性」の要件**として 「ポジティブな波及効果を持つこと」 を取り上げる考え方も
＝波及妥当性 washback validity
  - ・評価の学習規定性 学習への方向づけの期待
  - ・大規模テスト＝教育利用 （授業の教材などとして）
- ➢ **ネガティブな波及効果** → テスト得点の妥当性は流動的
  - ・テストワイズネス test-wiseness 選択肢の長さなどの手がかりを利用する力
  - ・テストスキル test-taking skill 適切な時間配分など自分の力を発揮するためのスキル

Cf. 村山航（2006）．テスト形式が学習方略に与える影響とそのプロセスの解明 東京大学教育学研究科博士論文

12

（4）教育へのインパクト

★ ハイステイクスな大規模試験の波及効果については、 負の効果もあり得ることから、慎重な検討が事前に求められる。

★ 波及効果は、 さまざまな要因（個人差・科目差・試験枠組など）によって、 そのあり方が異なってくることに留意すべきである。

★何らかの波及効果が受験生に及ぶことによって、テストの妥当性が変動することがある。測ろうと思っている特性とは別の要素が入り込んでくる。

★大規模試験の問題は、 試験という文脈で必要とされることと、大規模試験が教育に利用されるということを想定した工夫、 この両者のバランスをとることが求められる。 （*ex*. 図や写真などの挿入）

13

★2018年はムーミンのセンター試験だった！

大阪大学外国語学部スウェーデン語専攻

平成30年度大学入試センター試験　地理歴史（地理B）第5問問4への見解

高橋美恵子、吉谷大輔、富野能之の三名は、平成30年1月13日に実施されました大学入試センター試験地理歴史（地理B）の第5問問4につきまして、スウェーデン語に基づく研究教育を行っている大学に属する者の立場から、以下のような見解を公開します。なおここに掲載する見解は大阪大学としての見解ではありません。あくまでも上記三名の見解です。またこの見解は、地理学ではなくスウェーデン語に基づく北欧研究を行う者からみた意見です。私たちがこの見解を公開する意図は、大学入試センターならびに本問に関わられた方々を批判するものではありません。私どももまたセンター試験の末端にあってこれに関わる者です。立場の違いはあれセンター試験に等しく関わる者として、センターならびに本問に関わられた方々の努力に敬意を払うとともに、センター試験の社会的価値を維持することに、私どもの研究で得られた知見を参考意見のひとつとして活かして頂きたいとの思いから、この見解を公開します。

問 4　ヨシエさんは，3か国の表を配置して，言語の違いに気づいた。そして，3か国の単語をモチーフにしたアニメーションが日本のテレビで放映されていたことを知り，3か国の文化の共通性と言語の違いを調べた。次の図5中のタとチはノルウェーとフィンランドを舞台にしたアニメーション，AとBはノルウェー語とフィンランド語のいずれかを示したものである。フィンランドに関するアニメーションと言語との正しい組合せを，下の①～④のうちから一つ選べ。 28

スウェーデンを舞台にしたアニメーション　スウェーデン語

アニメーション　言語

タ「ムーミン」　A　Hva koster det?

チ「小さなバイキング ビッケ」　B　Hvor mye koster det?

図 5

	①	②	③	④
アニメーション	タ	タ	チ	チ
言語	A	B	A	B

14

第５問　東京在住の高校生のヨシエさんは，次の図１に示すルートのように，デンマークを経由して，ノルウェー，スウェーデン，フィンランドを旅行した。そして，夏休みの宿題として３か国を比較したレポートを作成した。このレポートに関する下の問い（問１～５）に答えよ。（配点　14）

図 1

Wikipedia　小さなバイキング

出題内容に対する見解　朝日新聞は大学入試センターに「小さなバイキングビッケの舞台はノルウェーと言えるのか，真相には主人公たちが住む村はスウェーデンのフラーケ地方にあるという記述がある」という内容の質問を行った。」

一　それに対し，大学入試センターは「バイキングの表記から海が結氷せず，海上活動が盛んだったノルウェーやスウェーデンを含むスカンディナビア半島の沿岸や周辺海域が想起されます。バイキングは現在のスウェーデンを含む地域においても活動していましたが，設問の設計上スウェーデンはすでに例示しており，また，フィンランドとは関連が薄いことから，ノルウェーが類推されます。」と回答した。センターの回答では，アニメの舞台にふれていないため，改めて朝日新聞が電話取材をしたところ，センターは「舞台がノルウェーであることを明確に示す出典は確認できていません。その意味では『ノルウェーを舞台にしたアニメーション』との記述は，厳密な意味で正確なものではありません」と答えた。ただ，センターは「知識・思考力を問う設問として支障はなかった」としており，出題ミスではないとしている。

一連の経緯を踏まえ，センターの大塚雄作・試験・研究統括官は「設問の趣旨を受験生に分かりやすく示すため，簡潔化した記述としたものであることをご理解いただきたい」とコメントした。

問 1　ヨシエさんは，3か国を旅行中に気象や気温の違いを体験し，3か国の気形と気候の特徴を次の図2のようにまとめた。図2中のア～ウは，3か国の標高200 m以下の面積*の割合と，ほぼ同じ緯度にある首市（ベルゲン，ストックホルム，ヘルシンキ）における最寒月と最暖月の月平均気温を示している。国名とア～ウとの正しい組合せを，次ページの①～⑥のうちから一つ選べ。 25

*高緯度を除く。

図 2

気象庁の資料などにより作成。

15

2. 記述式問題とその採点

（1）教育文脈において共有される論点

（2）測定対象によって異なる採点基準

（3）採点者信頼性の問題

（4）採点コストと労力の問題

16

（1）教育文脈において共有される論点と記述式問題

ex. **2017年6月15日実施のフランス・バカロレア哲学試験**

バカロレア資格とはフランスの大学入試に相当する試験。毎年**6月中旬から<u>およそ一週間</u>**にわたって行われる。中でも初日の哲学試験は**<u>4時間にわたって</u>**一つの命題と格闘することを要求される。

（1）知るためには観察するだけで足りるか。
(Suffit-il d'observer pour connaître ?)

（2）権利を全て実行することは正当か。
(Tout ce que j'ai le droit de faire est-il juste ?)

(http://societas.blog.jp/1066467716　参照)

→ 解答には無限の多様性と広範囲のレベルが想定される
記述の論点は教育のプロセスにおいて共有されているべき

17

●『教育評価の基礎』記述式問題例

（京都大学全学共通教育・2012年度後期・大塚雄作担当授業・試験2013.1.22実施）

【問２５】**<u>授業においてグループ討論のテーマとしたように</u>**、今、中央教育審議会の大学教育部会では、入試のあり方、また、高校教育までの初中等教育と、大学教育とをどうつないでいくか（「高大接続」と呼ばれている）ということが議論されている。

そこでは、これからの教育では、「生涯を通じ不断に主体的に学び考える力、予想外の事態を自らの力で乗り越えることのできる力、グローバル化に対応し活力ある社会づくりに貢献することのできる力」などをもった人材を輩出することが教育に求められるとされている。

そのために、学校教育では、より具体的にどのような学力、どのような能力を身につける必要があるのか、そして、その教育に大きな影響を及ぼす大学入試をどのような形にしていけばよいのかということが課題とされている。

そこで、この課題に関して、あなた自身はどのように考えるか、以下の**<u>教育評価の用語を用いながら論じなさい</u>**。

<u>総括的評価</u>　　**<u>識別力</u>**
<u>妥当性</u>（日常的な意味での「妥当性」ではなく、測定・評価に関わる意味：validity）
<u>アカウンタビリティ</u>（**<u>説明責任</u>**でもよい：accountability）　　**<u>羅生門アプローチ</u>**

18

●採点の観点とルーブリック例

《問題寸評》　（試験翌週に学生への配付資料に採点基準等を掲載）

【問２５】 採点の基準（ある意味でのrubric）は、おおよそ、以下の通り。

①「取り上げた能力」および「それを育てる教育」が具体的に提示されているか： 4＝「能力」が的確に取り上げられ、その育成にふさわしい教育が具体的に提案されている。 3＝提案はあるが、その関連性や具体性などに十分とは言えない部分がある。 2＝提案はあるが関連性や具体性などが十分とは言えない。 1＝記述を試みているが、十分な提案にはなり得ていない。 0＝この点に関する記述がない。

②その教育の過程で行われるべき教育評価手法： 4＝①で取り上げた「教育」等において適切な評価手法が具体的に取り上げられている。 3＝提案はあるが、その教育へのつながりや具体性などに十分とは言えない部分がある。 2＝提案はあるがその教育へのつながりや具体性などの点で十分ではない。 1＝記述を試みているが、十分な提案にはなり得ていない。 0＝この点に関する記述がない。

③「取り上げた能力」を的確に評価する入試方法の提案： 4＝①で取り上げた「能力」等を的確に評価する入試方法が具体的に提案されている。 3＝提案はあるが、その「能力」への関連性や具体性などに十分とは言えない部分がある。 2＝提案はあるがその関連性や具体性などの点で十分ではない。 1＝記述を試みているが、十分な提案にはなり得ていない。 0＝この点に関する記述がない。

等々、**全10観点**により採点 → 10観点の合計点のα係数=0.859とまずまずの内部一貫性の程度を示す

19

●論述式問題の全体得点上位者の解答例とコメントの提示

◆今後の教育においてはとりわけ「生涯を通じ不断に主体的に学び考える力」が重要であると私は考える。なにかひとつでも複数でもかまわないが 自分が「これなら夢中になれる」という人生における研究対象があれば実りある有意義な生活を送ることができると思うからだ。さてそのような力を高校までにおいていかにして育んでいけばいいかということだが、やはり何をするにしても基本的な知識やスキルは必要不可欠である。それを身につけるためには目的をあらかじめ設定するような工学的アプローチではなく、多角的視点から評価していくために**羅生門アプローチ**を取り入れていく必要がある。一見**妥当性**の低そうな活動であっても別の視点から考えると非常に有用であったり学習者自身のモチベーション向上につながったりしていることもあるからだ。それを評価者が単体で評価するのは非常に困難なので、トライアンギュレーションも欠かせないところである。さてその評価をいかにして被評価者＝学習者）およびステークホルダーに伝達するかという**説明責任**の問題が出てくるわけであるが、ポートフォリオ・アセスメントとルーブリックの活用が有効ではないかと思う。ポートフォリオによって学習者の到達度やそこに到るまでのプロセスを評価者とともに自己評価することができるし、ルーブリックによって現段階で学習者に不足している事項を確認することもできる。あるいは学習共同体ということを考慮するならば、一斉テストを実施し、その相対評価の結果から学力の近い者同士でグループを組んでお互いの弱点を補完し合うのも良いかもしれない。このように目標準拠評価と集団準拠評価を上手に融合していくことが望ましいと思われる。そうして高校までに学習者が身につけた知識やスキルの**総括的評価**として大学入試は位置づけられるべきである。あるいは今後の大学での身の振り方を考えるためのステップという意味では診断的評価としての機能も備えているとも言える。そこで前述の力をどのようにして把握すればよいかであるが、基本的には現行のままで良いと思う。つまり第一段階として大学入試センター共通一次試験で基本的な知識と処理スピードおよび**識別力**を測定し、第二段階として大学の二次試験を実施する。いずれの試験も改善の余地は大いにあると思うが、教育現場のシステムの改善が優先されるべきだと思う。（979字）

→ 後半が弱く、「識別力」の言葉の使われ方は微妙だが、前半は、ポートフォリオ評価、ルーブリック、自己評価などの用語も活用され、的確に記述されている。高校教育の総括評価として、また、大学教育の診断的評価としての大学入試の位置づけ方もよい。高校までの教育や入試そのものの評価に関しては、十分な記述がないのが難点であるが、全体的にコンパクトに論が展開されており、**好感の持てる回答**である。

20

（2）受検者のレベルによって異なる採点基準

***ex.*「テスト得点の信頼性と妥当性について論じなさい」の採点基準案**

A＝信頼性・妥当性の意味がそれぞれ正確に把握されており、その検証方法についても適切に触れられている。また、信頼性と妥当性の関係性（**信頼性が高くないと高い妥当性は得られない**・**信頼性を高めることによって妥当性が低くなる場合がある**など）について的確に論じられている。

B＝信頼性・妥当性の意味とその検証方法については把握しているが、その両者の関係性については十分に論じられていない。あるいは、両者の関係性については触れられているが、検証方法について触れられていない。

C＝信頼性・妥当性の意味については正確に把握されているが、その検証法、及び、両者の関係性については触れられていない。

D＝信頼性・妥当性について両方の意味を正確に把握しているとはみなされない。

◆問題の難易度を下げるとすれば・・・

A′＝信頼性と妥当性の両方の意味が正確に把握されている。

B′＝信頼性と妥当性の一方の意味は正確に把握されている。もしくは、やや曖昧さがある。

C′＝信頼性と妥当性の意味が全く捉えられていないとは言えないが、把握していない可能性も窺える。

D′＝両方の語の意味がほとんど把握されていない。

21

（3）採点者信頼性の問題

➢**記述式採点は「採点者信頼性」を高く維持することが難しい**

一人の採点者内でも採点基準は変動しがち

採点者間の採点結果の変動はかなり大きい（右図参照→）

Cf. 池田 央（1992).『テストの科学』 日本文化科学社

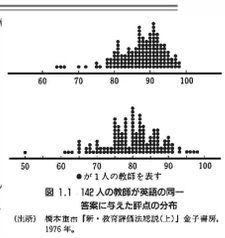


図 1.1　142 人の教師が英語の同一答案に与えた評点の分布

（出所）　橋本重治『新・教育評価法総説（上）』金子書房，1976 年。

➢海外の記述式を採用している試験では、アピール・センターを設置して、**記述式問題の採点に関わるクレームを受け付けている**場合もある

22

●記述式問題の得点の妥当性

◇条件付き記述式問題

制限字数、段落の分け方の指定、用いる用語の指定などにより、その条件ごとに採点を行い、それらの正答パターンに従って、点数や評定段階を割り振る方法

➢この方法によって、採点の安定性が確保され、得点の信頼性を担保することができる

➢しかし、得点の妥当性が犠牲になる場合がある

ex. 全国学力調査の記述式問題例→

★事前に十分な信頼性、妥当性に関わる検証研究の積み重ねの要

http://www.mext.go.jp/a_menu/shotou/gakuryoku-chousa/sonota/1347088.htm

問題番号		解答類型	正答
2	三	（正答の条件） 次の条件を満たして解答している。 ① **ア**に、**【雑誌の記事】**を読んで「宇宙エレベーター」について疑問に思ったことを一つ書いている。 ② **ア**に、「なぜ」、「どのような（に）」、「どのくらい」という言葉のいずれかを使って、二十字以上、四十字以内で書いている。 ③ **イ**に、必要な本の探し方を一つ書いている。 （正答例） ・**ア** 宇宙エレベーターは、なぜ長い期間にわたって繰り返し使うことが可能なのか。（37字） **イ** 図書の分類に従って、自然科学に関係する本が置いてある棚に行く。 **イ** 図書検索用のコンピュータに「宇宙エレベーター」と入力して検索する。 ・**ア** 宇宙エレベーターの実現には、どのような課題があるのか。（28字）	
	3	条件①、②を満たし、条件②を満たさないで解答しているもの （例） **ア** 宇宙エレベーターのケーブルは、地上のどこに向かって伸ばしていくのか。（35字） **イ** （略） ・**ア** 宇宙エレベーターに乗るための費用は幾らなのか。（24字）	誤答

23

（4）採点コストと労力の問題

★現段階では、大規模・共通一斉試験に記述式問題を導入する最大の問題点は採点の過程にあり、それにかかる時間、コスト、信頼性などの課題あり

★現行日程では、マークシート採点でさえ、マークの濃度や修正の仕方、あるいは消しゴムの滓などの要因によって、2度読み込んで同じマークと認定できない場合の確認作業など、現行のスケジュールでも精一杯というのが現実

★採点の信頼感を確保するために、民間が担当するにしても、大学入試センターが担当するにしても、高校教員、大学教員をいかに確保するかが問われることになると考えられる。→当事者意識を持つべし

24

3. 択一式問題に求められているもの

（1）個々の小問と試験得点との関係性
（2）当て推量をどう考えるのか
（3）マークシート方式問題の改善に求められるもの
（4）個別試験と共通試験の役割分担再考

25

●新テストに向けてのマークシート式問題の見直し

（2017.7.13公表『新テスト実施方針』より）

○マークシート式問題について、各教科・科目の特質や難易度を含む識別力の観点も踏まえつつ、思考力・判断力・表現力等を一層重視した作問への見直しを図るため、特に次のような点に留意して作問の工夫・改善に努める。

- 出題者が問題文で示した流れに沿って解答するだけでなく、**問題解決のプロセスを自ら選択しながら解答する部分が含まれる**ようにする
- **複数のテキストや資料を提示し、必要な情報を組み合わせ**思考・判断させる
- **分野の異なる複数の文章の深い内容を比較検討**させる
- 学んだ内容を**日常生活と結びつけて**考えさせる
- 他の教科・科目や社会との関わりを意識した内容を取り入れる
- **正解が一つに限られない問題**とする
- 選択式でありながら複数の段階にわたる判断を要する問題とする
- 正解を選択肢の中から選ばせるのではなく必要な数値や記号等をマークさせる

○**学習指導要領の趣旨・内容との連携をより的確に確保する**とともに、評価すべき能力や作問の構造を実際の作題に確実に反映するため、センターにおいては、高等学校関係者や、高等学校教育の実態をよく把握している大学教員等を積極的に作問委員として委嘱するなど、作問方針や体制の抜本的な見直しを図る。

26

●正解が一つに限られない問題

➢①～⑦のうち正しいものを**すべて選べ**

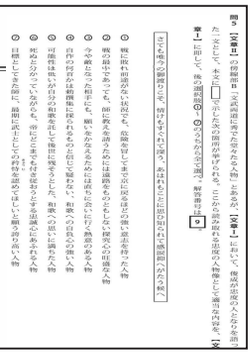
$=2^7=128$択

＝　○×の7つの小問

➢フィージビリティスタディの正答率

正答2.5%　無回答6.7%

→　難問・奇問化



27

【参考】池田央 (1992).『テストの科学―試験に関わるすべての人に』

➢「客観テストは“はい”か“いいえ”しか答えられないような無口な人から、その人が頭の中で考えている深い内容の話を聞き出そうとするようなもの･･･それには質問者の側でいろいろな角度から質問を発する必要がある･･･**客観テストでは、問いかけの質問数を多くして出力情報を増やした「細目積み上げ方式」と組み合わせて、そのカが発揮される･･･**」

（池田央・４章　客観テストの設計・p.84より）」

28

●細目積み上げ方式の考え方

（池田央『テストの科学』日本文化科学社より）

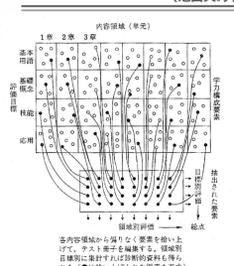


図 3.1　細目積み上げ方式の問題項目作成編集の考え方

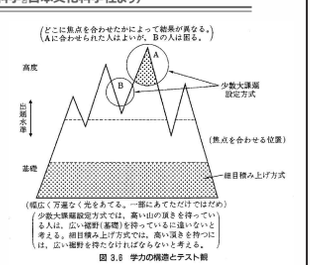


図 3.8　学力の構造とテスト観

29

◆解答形式は単純に

➢「ある問とある問と両方ができてはじめて正解にするとか、５枝選択のなかで特定の２枝を選んだ人のみに点を与える･･･数問分の選択枝を共通にして多くの選択肢から選ばせる･･･解答分析の結果はあまり思わしくない･･･**選択枝を複雑にすることは、その問で本当にみようとしているねらいとは別の要素が入り込んで得点の意味を複雑に**してしまう･･･　→　分割して問題数を増やす　＝　信頼性を高める」

➢「問題の難易は解答形式の複雑さよりも**出題の内容そのものにおくべき**であって、解答形式を複雑にすることによって問題を難しくするのは邪道･･･解答の仕方は画一的であっても、そこで要求される思考は、問題の内容によってさまざまに変えることができる･･･」

→　**選択枝の数を確認する**

→　**選択枝の長さを確認する**
ex. 幹に含められる場合は幹に入れ、枝は単純に

（池田央・４章　客観テストの設計・p.89～93より）」

30

（1）個々の小問と試験得点との関係性

- 個々の**小問の総体として試験得点**が求められる。試験得点が、いかに安定し、それが何を意味しているかを検討するのが、信頼性・妥当性検証の本来の観点。
- 個々の小問は、試験得点が、信頼性が高く、また、妥当性が高い得点となるように機能すべき。
- 項目分析（個々の小問の良し悪しに関する分析）の一つの指標として、**難易度と識別力の双方の観点**から、それぞれを反映する統計指標によって行われることがある。 右図参照 →

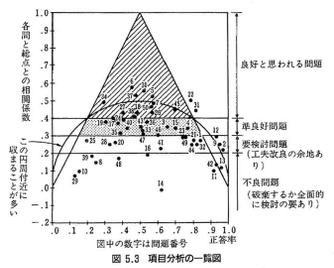


図 5.3 項目分析の一覧図

31

（2）当て推量をどう考えるのか

- 「選択枝数をできる限り多くしたいという心理の裏には、まぐれ当たりによる正答を少なくしたいという気持ちが働いている・・・」
  - → **問題数を増やす ＝ 総得点の信頼性が確保**されていればよい
  - → **難易度の高い問題**は **当て推量の解答と区別が難**しくなり **総得点の信頼性が下がる**
- 「選択枝の数が少なすぎるのもよくないし、多ければよいというものでもない・・・**枝の数を多くしようとすると、それだけ設問を用意するのが難しくなり、その結果、そこで狙っている事柄とは本質的に無関係な選択枝を入れたりすることになる**・・・」
- 「選択枝の本来のねらいは、ある一つの次元（枠組み）に関して拮抗するいくつかの比較材料を判断させて最適なものを選ばせようとするもの・・・判断次元に対して別の次元に属する"まよわし"を混入させると、解答者は観点を絞ることができず、いたずらに思考を混乱させて、本当にみたいと思っていた知識や能力が必ずしも表出されずに終わってしまう危険 → **妥当性が低下する原因**」
- 「択一式テストの**選択枝の個数は４つか５つが最も実用的**で、また理論的にも十分である・・・（**３つが最適であるという研究も**）・・・実質選択枝数が５を超えることはあまりない・・・」

（池田央・４章 客観テストの設計・p.85～87より）」

32

（3）マークシート方式問題の改善に求められるもの

① 客観テストは採点を客観的に行うように工夫されたテストであるが、解答者の出力情報が貧弱なため、**細目積み上げ方式**により効力を発揮する。

② 一つの客観テスト問題で、**設問から解答までの思考距離をあけない**ようにする。**全体を独立な小問に分解し、発問を多角的にして出力情報**を増やす。

③ **出題形式は、むしろ形式を統一**する。解答形式を複雑にすると、みたいと思っている能力要素以外の要素（不注意・不安傾向・思い違い等）を持ち込む危険性が高い。

④ **当て推量の心配は問題数を増やすことによって補う。選択枝を増やしたり、特別な工夫をすることは目的外の要因を取り込むことになりかねない。**

⑤ **問題数を減らしたり、一つの問題にあれこれ詰め込んだりすると、複数正解など、不適切な結果をもたらしやすい。**

⑥ **客観テストのよしあしを決める要素は、一問一問も大事だが、全問の構成内容や編集の仕方に依存するところが大きい。セット全体の問題構成に十分注意。**

（池田央・４章 客観テストの設計・p.98～99より）

33

（4）個別試験と共通試験の役割分担再考

●センター試験が果たしてきている共通試験としての役割

- **高校→入試→大学の高大接続の一体的流れの確保**
  - 大学「**が**」 大学入試センターと共同して行う 「**大学**」の試験
  - 大学の教員が 高校の学習指導要領に準拠しつつ 教科書なども精査して **大学教育の視点から**問題作成
- **難問奇問を排除した良質な問題の確保**
  - 問題作成＝第１委員会 ← 問題作成**OB**委員会によるチェック
    ＋**高校教員等**を中心とする点検協力者によるチェック
  - 試験実施後の問題公開による各方面からの問題照会
- **個別試験との組合せによる入試の個性化・多様化**
  - 小論文・面接 推薦入試・AO入試 帰国子女・社会人対象特別入試 *etc.*

34

（4）個別試験と共通試験の役割分担再考

●アドミッションポリシーに応じた個別試験の工夫

- **それぞれ独自の特性を評価する必要性**
  - ***ex.*** **数理的能力・言語表現力・芸術的能力・スポーツ的能力・語学力・・・*etc.***
  - それぞれの特性を信頼性・妥当性・等を十分に備えた入学試験方法の採用
  - そのために与えられたスケジュールのなかで最適な「サイズ」への絞り込みの要
- **多様な他者と協働する機会を教育において実現するために**
  - 多様なそれぞれの特性は、同じ入試方法では評価しきれない
  - それぞれの特性を的確に評価可能な**多様な入試枠組**を用意する要
- **選抜型競争入試から育成型入試への発想転換**
  - ボリュームゾーン以下の受験生の増大 → どう育成し、入学させ、どう教育していくか
  - この層に適切な共通テストを新たに導入していく必要はないのか？

35

●結 語

★「良問」は、問題自体が決めるものではなく、試験全体、あるいは、教育のプロセス全体の枠組において評価されるべき相対的なものである。

★いわゆる「学力の三要素」なるものを評価すべきということであるなら、まず、それに見合う試験の枠組とはどういうものであるかを、多層的・多面的に検討する必要がある。

★波及効果はやってみないとわからない点もあり、短絡的にそれだけを期待する試験の改変は、経済的・人的コストが大幅に増えるという状況の下では慎重に進めるべきである。

★試験においては、ただ実施するのみならず、妥当性等の検証のための追跡調査や、波及効果などを検討するフォローアップ調査を、本来は、組織的・体系的に導入していくべきである。教育実践のなかでは、それは実施可能性という点で難しいが、その視点は常に意識しておくことが望まれる。

★教育実践における評価の基本は、フィードバックにあるということ、教授者と学習者の相互作用を引き出すことにあるということに留意しておくことが望まれる。

了 36

高次の能力を捉えるための評価
ーパフォーマンス評価のデザインー

京都大学　高等教育研究開発推進センター　教授　松下　佳代

第23回FDフォーラム・第5分科会

2018.3.4@京都産業大学

高次の能力を捉えるための評価
ーパフォーマンス評価のデザインー

松下　佳代
京都大学・高等教育研究開発推進センター
matsushita.kayo.7r@kyoto-u.ac.jp

この分科会のねらい

- 「高次の能力を捉えるための評価」
  ～どのような評価がどのような能力を捉えることに適しているのかを課題づくりも含めて考える～
- 背後にある問題意識
  - 昨今、高次の（統合的な）能力の評価法に関する議論が多く行われている。
  - しかし、例えばルーブリックなど評価基準への注目は集まっているが、当該の能力を可視化するための課題の作成やその良し悪しの検討といったことに関する議論はまだ発展の途上にある。

3

高次の（統合的な）能力とは？

- 高次の（higher-order）
  ＊ブルーム・タキソノミーの上位3つの認知プロセス（＝分析・総合・評価）を指して使われることが多い
  ＝因果関係を分析したり、複数の関係や構造を総合したり、政策や価値判断を評価したりするような複雑な思考（田中編, 2005, p. 99）
  ＝〈暗記した知識の再生やなじみのある文脈での適用〉以上の認知プロセス
- 統合的な（integrative）
  ＝複数の単元や科目をまたがって、知識・技能を統合することを求められる

4

OUTLINE

- 評価の枠組み
- パフォーマンス評価の事例（1）
  ―京都大学初年次ゼミでのレポート評価―
- パフォーマンス評価の事例（2）
  ―新潟大学歯学部のPBL評価―
- パフォーマンス評価の事例（3）
  ―デジタル・リテラシーの評価―
- まとめ

5

評価の枠組み

6

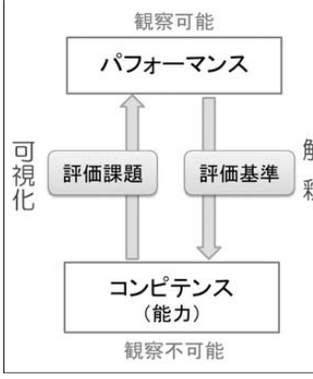
「ルーブリック評価」か、「パフォーマンス評価」か

- 「ルーブリック評価」？
  - performance assessment　391万ヒット
  - rubric（-based) assessment　4.8万（9千）ヒット　（Google検索）

  ➔「ルーブリック評価」は日本の、特に大学教育のジャーゴン
- パフォーマンス評価とルーブリックの関係
  - ルーブリックはパフォーマンス評価の評価基準の一つ
  - パフォーマンス評価と切り離したルーブリックの議論は意味がない

7

評価の枠組み



- どんな能力も、それ自体は観察不可能。
- そこで、能力を、何らかの 評価課題を通じて可視化させ、観察可能なパフォーマンスにする。
- そして、そのパフォーマンスを、評価基準を介して解釈することで、パフォーマンスの背後にある能力を推論する。

＊コンピテンス≒コンピテンシー

「評価とは、学生が知っていることについての合理的な推論を行うために、学生の行動の観察やデータの産出を行うよう作られたツールである」（NRC, 2001, p.47）

8

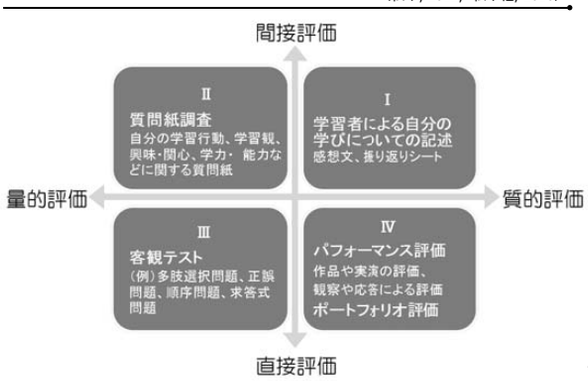
評価で考えるべきこと

- ①コンピテンス（コンピテンシー）
  - どんな能力を評価するか？（←どんな能力を目標とするか）
- ②パフォーマンス
  - どんなパフォーマンスを求めるか？
    （授業中の応答・言動、テスト、作品（プロダクト）、実演など）
- ③可視化
  - どんな評価課題を設定するか？
- ④解釈
  - どんな評価基準を設定するか？
- ⑤評価主体
  - 誰が評価するのか？（学生自身、教師、上級学校、テスト機関、実習先など）

9

学習評価の4つのタイプ

（松下, 2012; 松下他, 2016）



10

間接評価と直接評価

（Banta & Palomba, 2015）

- 間接評価
  - 学生の学習行動や学習についての自己報告を通じて、学習成果を間接的に評価
  - 質問紙やミニッツペーパーなど

  何ができると思っているか？
- 直接評価
  - 学生の知識や能力の表出を通じて、学習成果を直接的に評価
  - 客観テストやパフォーマンス評価など

  何ができるか？

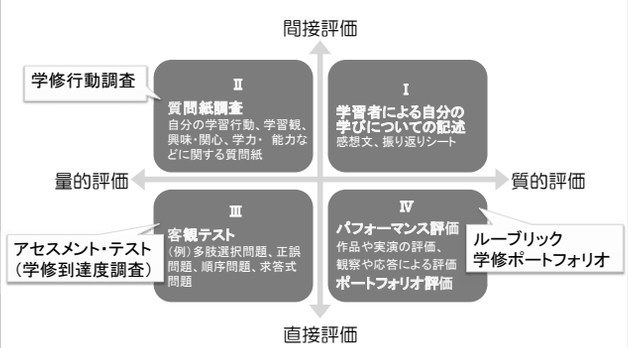
11

量的評価と質的評価

（松下, 2017）

	量的評価	質的評価
学問的基盤	心理測定学	解釈学、構成主義的学習論など
評価データ	量的データ	質的データ
評価対象	集団または個人	個人
評価目的	選抜、組織的な教育改善、アカウンタビリティなど	学習や指導の改善など
評価課題	細かく分割された問題 文脈独立的	複合的な課題 文脈依存的
評価基準	客観性を重視	間主観性を重視
評価結果	数値	文章や数値
評価機能	主に総括的評価	主に形成的評価
評価方法	客観テスト・標準テスト、質問紙調査など	パフォーマンス評価・ポートフォリオ評価、ミニッツペーパーなど

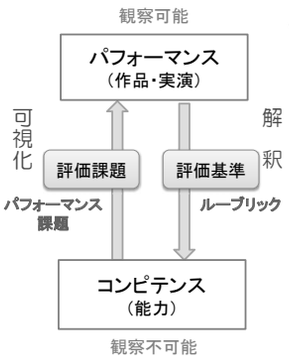
12



cf. アセスメント・ポリシー（「質的転換答申」2012.8）
「特に、成果の評価に当たっては、学修時間の把握といった学修行動調査やアセスメント・テスト（学修到達度調査）、ルーブリック、学修ポートフォリオ等、どのような具体的な測定手法を用いたかを併せて明確にする。」

13

パフォーマンス評価（タイプIV）



- 2つのアプローチ（Messick, 1994）

【1】パフォーマンスそのものに焦点
（例）芸術コンテスト、スポーツ競技

【2】パフォーマンスを可能にしているコンピテンス（能力）に焦点
＊特に教育ではこの視点が重要

14

パフォーマンス課題

- パフォーマンス課題 (performance task)
  - 学習者のパフォーマンスを評価するためにデザインされた課題
  - リアルな文脈のもとで、さまざまな知識やスキルを総合して使いこなすことを求めるような課題

「簡単にパフォーマンス評価の意図をいえば、テストの内容を、基準となるパフォーマンスで示される批判的な思考や知識の総合を求めるものにしていこうとするところにある」（ギップス, 2001, p.16）

＝高次の（統合的）能力

- 大学教育の現状
  - パフォーマンス課題は多く使われている
    - 実技（医療・教員養成など）、演奏や創作物（芸術分野）、製作物（PBLなど）、レポート・卒業論文、プレゼン・口頭試問など
  - しかし、評価基準については、ほとんど主観にゆだねられていた

15

ルーブリック

- ルーブリックとは

＝パフォーマンスを能力の表れとして解釈し、その質（よさ・見事さ・美しさ）を段階的・多面的に評価するための評価基準表

- 「専門家の鑑識眼」を明示化し、共有できるようにするツール
- パフォーマンスの質を量的表現に変換する（質的評価と量的評価をつなぐ）働きも

	レベル4	レベル3	レベル2	レベル1
観点1	……	……	……	……
観点2	……	……	……	……
観点3	……	……	……	……
観点4	……	……	……	……

レベル
観点（規準）
記述語（descriptor）

16

- ルーブリックのタイプ
  - （a）構造
    - 観点を複数設定して分析的に評価する「分析的（観点別）ルーブリック」
    - 観点を分けずに全体的に評価する「全体的（包括的）ルーブリック」
  - （b）スコープ（適用範囲）
    - ある領域で一般的に適用できる「一般的ルーブリック」
    - 当該課題だけに適用される「課題特殊的ルーブリック」
  - （c）スパン（対象期間）
    - 複数年にまたがって使われる「長期的ルーブリック」
    - 採点のためにスナップショット的に使われる「採点用ルーブリック」

17

パフォーマンス評価の事例（1）

ー京都大学初年次ゼミでのレポート評価ー

18

私自身の授業での取り組み

- 全学共通科目「学力・学校・社会」（初年次セミナー）

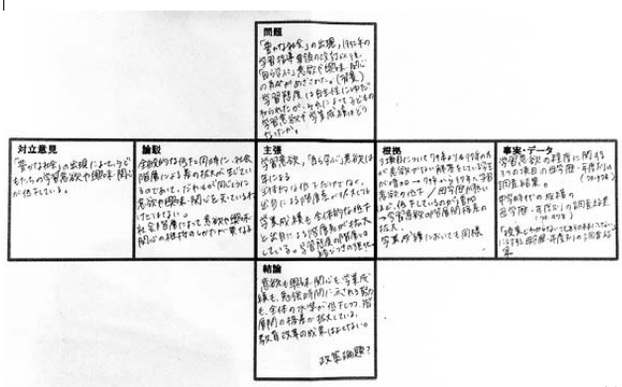
【目標】
- 内容に関する目標：学力・能力を軸として、学校（大学を含む）と社会の関係について考え、自分なりの見方をもつ。
- 能力に関する目標：批判的に読み、議論し、書くことができるようになる。

【授業構成】
- 前半： 講義とディスカッション
- 後半： プレゼンテーションとピア・レスポンス

【評価】
- 最終レポート課題：「授業に関連するテーマを自分で設定し、モデルにそって論を構成する。プレゼンテーションの内容について、フロアからの意見・感想・質問をふまえて修正・補足し、文章化する。」
- ルーブリックの活用 （新潟大学歯学部で共同開発したもの）

19

■論証モデルを使って苅谷剛彦氏の議論を分析する（第4回）



21

■モデルと連動したルーブリック

ILASセミナー「学力・学校・社会」

レポート評価基準

観点	問題解決	論理的思考				文章表現
	背景と問題	主張と結論	根拠と事実・データ	対立意見の検討	全体構成	表現ルール
観点の説明	与えられたテーマから自分で問題を設定する。（問題については、事実論題、価値論題、政策論題が区別されている。）	設定した問題に対し、展開してきた自分の主張を関連づけながら、結論を導く。	自分の主張を支える根拠を述べ、根拠の真実性を立証するため事実・データを明らかにする。	自分の主張と対立する意見を取りとげ、それに対して論駁（問題点の指摘）を行う。	問題の設定から結論にいたる過程を論理的に組み立て、表現する。	研究レポートとしてのルール・規範を守り、適した文章と言い回しを用いてレポートを作成する。
レベル3	与えられたテーマから問題を設定し、論理の整合性、その問題を取りとげた理由や背景について述べている。	設定した問題に対し、展開してきた自分の主張を関連づけながら、結論を導いている。結論は一般論にとどまらず、独自性を有している。	自分の主張の根拠が述べられており、かつ根拠の真実性を立証する信頼できる複数の事実・データが示されている。	自分の主張と対立するいくつかの意見を取りとげ、それらすべてに対して論駁（問題点の指摘）を行っている。	問題の設定から結論にいたる論理的な組み立て、記述の順序、パラグラフの接続が整っている。	・制限を行っていない、引用部分と自分の文章の区別を明示し、引用部分については、出所を記載している。 ・研究レポートとして適した文章と言い回しを用いてレポートを書いている。 ・参考文献・参考サイトを明示している。
レベル2	与えられたテーマから問題を設定し、その問題を取りとげた理由や背景について述べている。	設定した問題に対し、展開してきた自分の主張を関連づけながら、結論を導いている。	自分の主張の根拠が述べられており、かつ根拠の真実性を立証する信頼できる事実・データが少なくとも一つ示されている。	自分の主張と対立する少なくとも一つの意見を取りとげ、それに対して論駁（問題点の指摘）を行っている。	問題の設定から結論にいたる論理的な組み立て、記述の順序、パラグラフの接続がおおむね整っている。	＊3つの条件をすべて満たしているものをレベル3、2つ満たしているものをレベル2、1つしか満たしていないものをレベル1とする。
レベル1	与えられたテーマから問題を設定しているが、その問題を取りとげた理由や背景の内容が不十分である。	結論は述べられているが、展開してきた自分の主張とのつながりが不十分である。	自分の主張の根拠は述べられているが、根拠の真実性を立証する信頼できる事実・データが明らかにされていない。	自分の主張と対立する意見を取りとげているが、それに対して論駁（問題点の指摘）がなされていない。	問題の設定から結論にいたるアウトラインはたどれるが、記述の順序やパラグラフの接続に難点のある箇所が散見される。	
レベル0	レベル1を満たしていない					

22

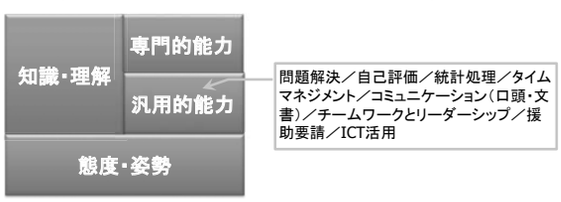
パフォーマンス評価の事例（2）
ー新潟大学歯学部のPBL評価ー

新潟大学歯学部小野和宏教授との共同研究

23

どんな能力を身につけさせるか

- 卒業時に獲得が期待される学習成果
  ＝“新潟大学歯学部版学士力”（小野・大内・前田, 2011）



⇨ 高次の統合的な能力にみあった評価が必要
＝能力目標と評価の整合性（alignment）

24

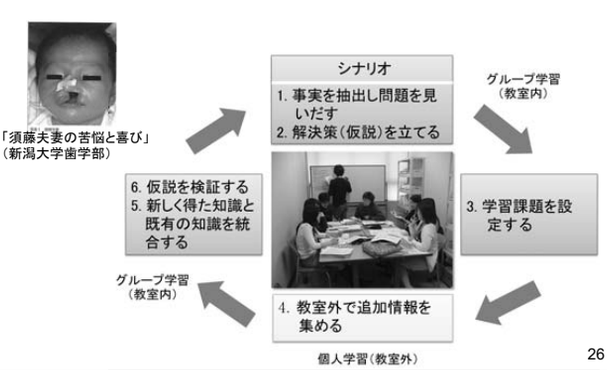
PBLのパフォーマンス評価に向けて

- 新潟大学歯学部でのPBLテュートリアル
  - 2004年度より実施
  - テューターの支援の下での7～8人のグループによる問題解決学習
- PBLテュートリアルでの学びをどう評価するか？
  - 筆記テストや卒業生調査では形成しようとする能力*が評価できない（*問題解決能力、自己学習能力）
  - ⇩
  - パフォーマンス評価の開発へ
    - PBLの評価として知られているトリプルジャンプ（Mtshali & Middleton, 2011）の改訂

25

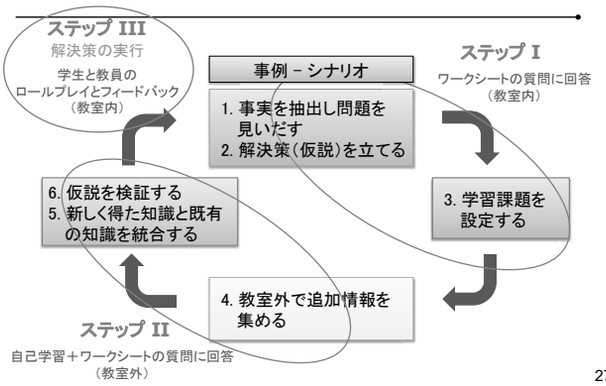
- PBL（Problem-Based Learning）のプロセス

（小野他, 2011; 小野他, 2015）



26

「改良版トリプルジャンプ」の開発



27

パフォーマンス課題のシナリオ

「わたし、困っています」

あなたは新潟大学医歯学総合病院の歯科衛生士です。
今日は担当患者：高橋勇蔵（67 歳・男性）の2回目の診療日です。高橋勇蔵は中等度の歯周病があり、初回は歯周検査と病状説明を行いました。

あなた：高橋さん、お口の具合はいかがですか。前回、タバコをやめるようお話しましたが、禁煙されましたか。
高 橋：してないよ。私はね、タバコをやめるくらいなら死んだ方がましだと思っているよ。この前、国から送られてきたアンケートにも、生きがいはタバコを吸うこと、と書いたくらいだ。相変わらず1日 40 本は吸っている。歯科に来て、なんでタバコをやめるよう言われなきゃならんのだね。
あなた：でも、高橋さんは糖尿病もあるし、やめた方がいいと思いますが‥。
高 橋：糖尿病は関係なかろう。ここは歯科だろう。おやおや、内科と間違えたかな。
あなた：歯科ですけど‥。とにかく、前回言ったことと同じことを言いますが、まずはタバコをやめてください、いいですね。
高 橋：ああ、わかった。あんたはタバコが嫌いだな。

28

改良版トリプルジャンプの構造

ステップ I
（PBLの学習プロセス1～3）
教室でワークシートに記述させる（60分間）

ステップ II
（PBLの学習プロセス4～6）
授業時間外にワークシートに記述させる（1週間）

ステップ III
（解決策の実行）
教員を相手に、ロールプレイさせ、
その結果をフィードバックする（15分間）

ルーブリック（学生に提示）により評価

ルーブリック（学生に非提示）により評価

29

新たに加えたステップIII

- ロールプレイ
  - 学生は、模擬患者役の教員とやりとりしながら禁煙指導のロールプレイ
  - ロールプレイを観察しながら、3人の教員が、ルーブリックにそって評価（約7分）
- フィードバック
  - ロールプレイ終了後、教員からすぐに、結果（改善点やアドバイスなど）をフィードバック （約8分）

＝学期最後の総括的評価だが、
形成的評価としての機能を強化

30

ルーブリック （小野他, 2015）

観点	問題発見～最終解決策の提案（ステップI・II）					
	1. 問題発見	2. 解決策の着想	3. 学習課題の設定	4. 学習結果とリソース	5. 解決策の検討	6. 最終解決策の提案
観点の説明	シナリオの事実から、問題を見いだす。	解決の目標を定め、いくつかの解決策を立案する。	問題の解決に必要な学習課題を設定する。	信頼できるリソースから、学習課題を調査する。	解決策の有効性や実行可能性を検討する。	問題に対して最終的な解決策を提案する。
レベル3						
レベル2						
レベル1						
レベル0						

観点	解決策の実行（ステップIII）			
	7-1. 追加情報の収集	7-2. 情報の統合	7-3. 共感的態度	7-4. コミュニケーション
観点の説明	禁煙を働きかける上で必要となる追加情報を患者とのやりとりを通じて収集する。	禁煙を働きかける上で有用な情報を結びつけて理解する。	患者の考えや価値観に配慮して禁煙を働きかける。	自分の考えを患者にわかりやすく説明する。
レベル3				
レベル2				
レベル1				
レベル0				

長期的ルーブリックとして開発

学生の意見・感想

- 「ステップIIIを行うことで、PBLが将来現場に出た時に役立つものになるのだと分かりました。」
- 「トリプルジャンプを行ってみて、現在のPBLよりも終わった後の手応えがとてもありました。トリプルジャンプは、調べ学習を行って分かった事柄を患者にどのように説明したらよいのか考えなくてはいけませんでした。ただ調べて終わりではなく、次のステップがトリプルジャンプにはあるので、自分の勉強した内容を自分の頭できちんと整理しなければなりません。よって、調べたことが自分のものになり、ためになりました。」
- 「これまでのPBLと違い、すべての作業を自分一人で行うので、大変さは感じました。調べきれなかったところは他の人が調べてきてくれるだろうという期待ができないので、今までで一番学習したように思います。大変さはあるのですが、すべて自分の責任になってくるので、モチベーションがあがって頑張れたのかもしれません。」
- 「やったらやった分だけ身になるような気がします。また、自分がどれくらい成長できるのかが分かるので、ただ学習しているだけよりも、もしかしたら手ごたえがあるかもしれません。実際一人ずつ評価コメントしなくてはならない先生方には大きな負担になるかもしれないのですが。」

32

改良版トリプルジャンプ（MTJ）の特徴

- 一人PBL
  - デメリット･･･「協同的に取り組む能力」は評価できない
  - メリット･････将来出会う現実の状況にはより近い
- PBLを補強
  - 通常のPBLサイクルは「解決策の検討」までで、「解決策の実行」「結果の評価」が欠落　➔　ステップIIIで補強
- 学習としての評価
  - 学生にとって、単なる〈学習の評価（assessment of learning）〉ではなく、〈学習としての評価（assessment as learning）〉になっている

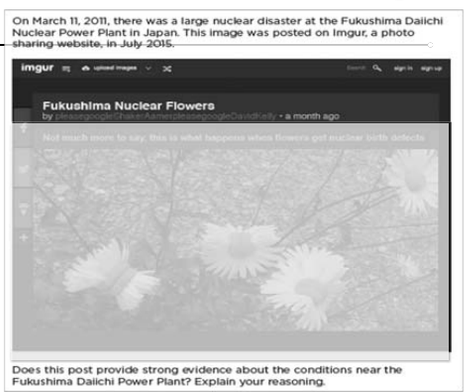
33

パフォーマンス評価の事例（3）
ーデジタル・リテラシーの評価ー

34

デジタル・リテラシー（civic online reasoning）の評価

- 評価課題



＊この課題は高校生対象　（Wineburg et al., 2016, p. 16）

35

- 評価基準（ルーブリック）
  ＊全体的・課題特殊的ルーブリック

MASTERY	生徒は、この投稿は強いエビデンスにならないと論じ、投稿の出所（例えば、投稿者についてわれわれは何も知らない）や写真の出所（この写真がどこで撮られたかわからない）などの疑問を呈する。	結論○ 理由○
EMERGING	生徒は、この投稿は強いエビデンスにならないと論じるが、その説明は、投稿や写真の出所を考慮したものではないか、あるいは、説明が途中で終わっている。	結論○ 理由△
BEGINNING	生徒は、この投稿は強いエビデンスになると論じる。または、不正確な（あるいは一貫性のない）推論を用いる。	結論× 理由×

（Wineburg et al., 2016, p. 18を一部改変）

36

デジタル・リテラシーの評価

- 先行研究
  - Stanford History Education Group（SHEG）の実践研究
    （https://sheg.stanford.edu/civic-online-reasoning）
- デジタル・リテラシー（Civic Online Reasoning）とは
  ＝オンライン環境にあふれる情報の信頼性を判断する能力
  ＝ポスト真実の時代を生きる市民に求められる能力
  - 3つのコンピテンシーからなる

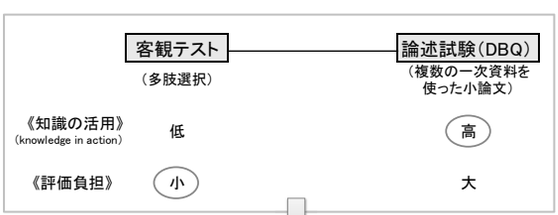
①誰が提示された情報の背後にいるのかを同定する
②提示されたエビデンスを評価する
③他の情報源が言っていることを調べる（lateral reading）

＊Fukushimaの問題は①②のコンピテンシーを評価

37

SHEGの評価開発の特徴

- 従来の評価：2つの極



- 〈知識の活用を把握することができ、しかも評価負担が小さい評価〉の開発

38

日本の大学でのデジタル・リテラシー評価

- 京都大学松下ゼミ「高等教育方法演習」で開発
  - 第24回大学教育研究フォーラム（2018）
    - 長沼祥太郎・杉山芳生・澁川幸加・浅川裕子・Jeong Hanmo・土岐智賀子・山田 勉・松下佳代「実行可能性を考慮したデジタル・リテラシー評価の開発」
    - 飯尾 健・香西佳美・溝口 侑・大森俊典・渡邊智也・平山朋子・小山理子・松下佳代「メディア情報リテラシーのパフォーマンス評価の開発」
  - 課題例「高度外国人材にとって日本は魅力的な国である」（長沼他, 2018）
    - 3つのコンピテンシー（〈1〉情報源の同定、〈2〉エビデンスの評価、〈3〉他の情報源との照合）を評価
    - 分析的・課題特殊的ルーブリックを使って自己評価・相互評価

39

- 課題例（〈1〉情報源の同定、〈2〉エビデンスの評価、〈3〉lateral reading［他の情報源との照合］）

40

41

- ルーブリック（〈1〉情報源の同定、〈2〉エビデンスの評価）
  ＊〈3〉lateral readingはレベル4に包含されている

	対象となる問題	4点	3点
【誰が情報の背後にいるのかを同定する】	①～③	記事の執筆者情報に関する箇所に印をつけており，かつ，外部のwebサイトから，記事作成者の情報を集めようとしている（〈3〉）	記事の執筆者情報に関する箇所に印をつけており，かつ，その人がどういった人なのかに関する記述がある
〈1〉 例		「この記事作成者が博士号を持っているという情報は見つけられない」等	「博士号を持つ人が書いているので，信頼していい」 「身元があやしい」等
【証拠や根拠の質を評価する】	①～③	記事の証拠や根拠に関する箇所に印をつけており，かつ，外部のwebサイトから，記事上の証拠や根拠の妥当性を検討している（〈2〉）	記事の証拠や根拠に関する箇所に印をつけており，かつ，記事の証拠や根拠の質の悪さに関する適切な記述がある
〈2〉 例		「他のWebサイトでは，高度外国人材にとっての魅力度ランキングで日本が最下位であると報告されているため」 「少子高齢化のグラフの引用元をたどったが，正しかった」等	「サンプル数が少ない」「調査対象が誰かが書かれていないので怪しい」「記事作成者の友人だと，都合のいいことしか言わないのではないか？」「結果を比較できていない」等
		2点	1点
		記事の執筆者情報に関する箇所に印をつけているが，その人がどういった人なのかに関する記述がない （誰がいてるだけ） あるいは 「？」など，具体的でない記述になっている	記事の執筆者情報に関する箇所に印をつけていない 同左
		記事の証拠や根拠に関する箇所に印をつけているが，それに関する記述がないもしくは，記述があっても内容が適切でない （誰がいてるだけ） あるいは 「何を言ってるかわからない」「うそっぽい」等	記事の証拠や根拠に関する箇所に印をつけていない 同左

42

まとめ

- パフォーマンス評価のデザインのポイント
  - ①コンピテンス（コンピテンシー）
    - どんな能力を評価するか？ ➔ 高次の統合的な能力であることが多い
  - ②パフォーマンス
    - どんなパフォーマンスを求めるか？ ➔ 作品（プロダクト）、実演など
  - ③可視化
    - どんな評価課題を設定するか？ ➔ 構成概念（内容知識、高次の統合的な能力）を反映した真正性の高い課題
  - ④解釈
    - どんな評価基準を設定するか？ ➔ ルーブリックの形式を取ることが多い
  - ⑤評価主体
    - 誰が評価するのか？ ➔ 教師、学生（自己評価、相互評価）など

44

- 3事例の特徴

	事例1（レポート）	事例2（PBL）	事例3（デジタル・リテラシー）
①コンピテンス	批判的思考力（論証モデルの活用）	問題解決能力、コミュニケーション	デジタル・リテラシー（Civic Online Reasoning）
②パフォーマンス	レポート	・ワークシート ・ロールプレイ	選択肢 ＋理由の説明
③可視化（評価課題）	レポート課題（プレゼンとピア・レスポンスをふまえて）	シナリオ課題（ストーリー性と真正性）	紙とウェブ検索
④解釈（評価基準）	ルーブリック（一般的）	2種類のルーブリック（一般的、課題特殊的）	ルーブリック（課題特殊的）
⑤評価主体	教員	教員（学生の自己評価も）	学生（自己評価・相互評価）

45

文 献

- Banta, T., & Palomba, C. (2015). *Assessment essentials: Planning, implementing, and improving assessment in higher education* (2nd ed.). San Francisco: Jossey Bass.
- Earl, L. M. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin Press.
- Gipps, C. V. (1994). *Beyond testing: Toward a theory of educational assessment*. Falmer Press. ギップス, C. V. (2001).『新しい評価を求めて―テスト教育の終焉―』(鈴木秀幸訳)論創社.
- 牧野由香里 (2008).『「議論」のデザイン』ひつじ書房.
- 松下佳代 (2007).『パフォーマンス評価』日本標準.
- 松下佳代 (2012).「パフォーマンス評価による学習の質の評価―学習評価の構図の分析にもとづいて―」『京都大学高等教育研究』第18号, 75-114.
- 松下佳代 (2016).「アクティブラーニングをどう評価するか」松下佳代・石井英真(編)『アクティブラーニングの評価』東信堂, 3-25.
- 松下佳代・小野和宏・高橋雄介 (2013).「レポート評価におけるルーブリックの開発とその信頼性の検討」『大学教育学会誌』第35巻第1号, 107-115.
- McGrew, S., Ortega, T., Breakstone, J., & Wineburg, S. (Fall 2017). The challenge that's bigger than fake news: Civic online reasoning in a social media environment. *American Educator, 41*, 4-9.
- Messick, S. (1994). The psychology of educational measurement. *Journal of Educational Measurement, 21*(3), 215-237.
- Mtshali, N. G., & Middleton, L. (2011). The triple jump assessment: Aligning learning and assessment. In T. Barrett & S. Moore (Eds.). *New approaches to problem-based learning: Revitalising your practice in higher education*. New York: Routledge.

- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. Pelligrino, J., Chudowsky, N., & Glaser, R., editors. Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- 小野和宏・大内章嗣・前田健康 (2011).「学習者主体PBLカリキュラムの構築―新潟大学歯学部口腔生命福祉学科7年のあゆみ―」『新潟歯学会誌』41(1), 1-12.
- 小野和宏・松下佳代・斎藤有吾 (2014).「PBLにおける問題解決能力の直接評価―改良版トリプルジャンプの試み―」『大学教育学会誌』36巻1号, 123-132.
- 小野和宏・松下佳代 (2015).「教室と現場をつなぐPBL―学習としての評価を中心に―」松下佳代・京都大学高等教育研究開発推進センター編『ディープ・アクティブラーニング―大学授業を深化させるために―』勁草書房, 215-240.
- 小野和宏・松下佳代 (2016).「初年次教育におけるレポート評価」松下佳代・石井英真編『アクティブラーニングの評価』東信堂, 26-43.
- 田中耕治編 (2005).『よくわかる教育評価』ミネルヴァ書房.
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press. トゥールミン, S. (2011).『議論の技法―トゥールミンモデルの原点―』(戸田山和久・福澤一吉訳)東京図書.
- Wiggins, G. & McTighe, J. (2005). *Understanding by design* (Expanded 2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development. ウィギンズ, G. & マクタイ, J. (2012).『理解をもたらすカリキュラム設計―「逆向き設計」の理論と方法―』(西岡加名恵訳)日本標準.
- Wineburg, S., McGrew, S., Breakstone, J., & Ortega, T. (2016). *Evaluating information: The cornerstone of civic online reasoning*. (http://purl.stanford.edu/fv751yt5934) (2017.8.27アクセス)