教育プログラム評価とは

大西弘高
医学教育国際協力研究センター

東京大学医学教育
國際協力研究センター

医学教育専門家の仕事

何らかの改革に携わり
その評価をする

→では，その評価はどうすべきなのか

→現在なされている評価は妥当か

例：新医師臨床研修制度

医学書院

サイト全体

検索

医学 | 看護 | コメディカル

HOME ＞ 週刊医学界新聞 ＞ 第2768号 2008年02月11日

第2768号

- 〔対談〕臨床研修制度の評価と展望（福井次矢，宮嵜雅則）
- 〔寄稿〕第2回EBM国際大会奮闘記（西崎祐史）
- 女性のためのキャリアデザインセミナー開催
- 〔連載〕名郷直樹の研修センター長日記（49）受け継がれること
- 〔連載〕生身の患者と仮面の医療者（終了）（11）内面化する仮面
- 〔連載〕レジデントのための栄養塾（終了）（7）頭に入れておきたい強制栄養の重篤な合併症
- 〔連載〕臨床医学航海術（25）医学生へのアドバイス（9）
- 〔連載〕はじめての救急研修（終了）（21）尿路感染では単純性か複雑性かを考慮！
- MEDICAL LIBRARY 書評・新刊案内

バックナンバー一覧

連載一覧

第2768号 2008年2月11日

【対談】

臨床研修制度の評価と展望

福井 次矢氏（聖路加国際病院・院長）
宮嵜 雅則氏（厚生労働省医政局医事課／医師臨床研修推進室・室長）

東京大学医学教育国際協力研究センター

――まず福井先生から，研修医の基本的臨床能力の修得状況に関する調査について，その概要をご紹介ください。

福井　私たちは20年近く前から研修医の臨床能力を調査してきました。その一環として，2003年3月に旧制度下の2年次研修医（n＝2474人）を対象にアンケート調査を行っています。そして，ほぼ同じ内容のアンケート調査を，2006年3月の2年次研修医（n＝1166人），つまり新しい制度の1期生が2年間の研修をほぼ終えた時点で行って，その結果を2003年3月に行った調査の結果と比較しました。

結論を申し上げますと，研修医による自己評価では，旧制度下の研修医（2003年3月の2年次研修医）に比べて，新制度下の研修医（2006年3月の2年次研修医）は，幅広い臨床能力の修得状況が著しく向上していることがわかりました（表）。特に大学病院の研修医はほとんどの項目で改善していて，伸び率が50％以上の項目も多数ありました。旧制度下では研修病院の研修医の臨床能力のほうが大学病院の研修医の臨床能力よりもかなり高いという結果でしたが，新制度下では両者の差がほとんど認められませんでした。

表　二値化(できるvsできない)のできる割合

調査項目ラベル	質問内容	新制度導入前 %	新制度導入後 %
a. 基礎的な臨床知識・技能			
細菌培養	グラム染色を行い，結果の解釈ができる	31.37	54.24
術後合併症	術後起こりうる合併症及び異常に対して基本的な対処ができる	52.81	74.59
輸液	輸液の種類と適応を挙げ，輸液の量を決定できる	77.88	87.45
創傷	傷病の基本的処置として，デブリードマンができる	48.37	69.05
症例呈示	カンファレンス等で簡潔に受持患者のプレゼンテーションができる	87.15	90.68
凝固検査	血液凝固機構に関する検査を指示し，結果を判定できる	85.99	92.72

恒常的な評価体制の構築へ

福井　昨年12月に，その審議会（医道審議会医師分科会医師臨床研修部会）の報告書がまとまったとのことですが，主だったところを教えていただけますか。

宮嵜　研修プログラムに関していえば，研修分野やその期間については現時点での変更はありません。ただ，原則として研修1年目は内科，外科，救急を，2年目に小児科，産婦人科，精神科，地域保健・医療をローテートすることになっているのですが，「実際の現場での運用や指導体制を考えると，ローテーションの仕方をもう少し柔軟にしてほしい」という意見が多くありました。こうした意見を踏まえ，1年目にも必修科（小児科，産婦人科，精神科，地域保健・医療）を3か月に限って可能とすることになりました。

福井　臨床研修の到達目標については，具体的な変更点はありますか。

宮嵜　いろいろご意見をいただきましたが，現時点では具体的な変更はありません。ただ，適時必要な改正ができるように恒常的な仕組みをつくることが提言されています。その仕組みづくりの議論を，早ければ年度内にも開始したいと思っております。

福井　昨年3月に，文科省の「医学教育の改善・充実に関する調査研究協力者会議」（座長＝高久史麿・自治医大学長）において報告書がとりまとめられましたね。あのなかで重要な提言のひとつが，「モデル・コア・カリキュラム改訂に関する恒常的な体制の構築」だと思います。評価と改善を恒常的に行う委員会が設置されたのは，大きな進歩です。

2008年9月10日

臨床研修制度のあり方等に関する検討会

臨床研修見直しで議論開始、年内に結論

http://headlines.yahoo.co.jp/hl?a=20080909-00000002-cbn-soci

医師不足を招いた一因とされる臨床研修制度を見直すため、厚生労働省と文部科学省は9月8日、「臨床研修制度のあり方等に関する検討会」の初会合を開き、意見交換した。「10年はかかる医学部の定員増よりも早く医師不足に対応できないかを議論するもの」(事務局)で、月1回のペースで行い、年内をめどに報告書をまとめる。検討結果が制度に反映されるのは早くとも2010年4月からだという。

舛添要一厚生労働相は「新研修制度は、プライマリーケアを育てるなど良い側面もある。問題の所在を見極めたい」としながら、「授業に魅力がなければ学生は来ない。医師を一人前に育てるにはどうするべきか、ここにメスを入れないといけない」と述べた。

ブログにて
地域医療日誌 by COMET

検討会の構成員の中でも、新臨床研修制度への意見は分かれた。
福井次矢・聖路加国際病院長は「臨床研修制度イコール医師不足ではない。どの専門分野にどれくらい足りないのか、そのデータが欲しい」と要求した。
小川彰・岩手医科大学長は「臨床研修制度が医師不足の要因となった。2年間の研修に専念するため、医療現場からは実際、1万5000-6000人のマンパワーが失われている」と訴えた一方で、西澤寛俊・西岡病院理事長は「専門教育を受けた若い先生が、地方に行きたがらない。臨床研修制度がなく、専門教育に重きを置いていれば、医療崩壊がさらに進んでいたかもしれない」と述べた。

どの専門分野にどれくらいの医師が必要か、データが重要です。しかし、この国の統計から試算できるものなのでしょうか。

東京大学医学教育国際協力研究センター

プログラム評価のスタンダード

（Joint Committee on Standards for Educational Evaluation, 1994）

1. プログラム評価の目的は何か
2. 評価に関する書面上の同意書は存在するか
3. 評価されるプログラムは明確に定義されているか
4. 評価者はプログラム評価をどのように計画するか
5. プログラムの“値打ち”を評価するにあたって、功績や価値の判断基準は明確か
6. 評価デザインはしっかりしたものか
7. プログラム評価を行うスタッフは充分であるか
8. プログラム評価のデザインや評価プロセスは改善が可能であるか
9. 評価結果は完全なもので、弁護・弁明ができるか
10. 評価結果の報告書に書かれた長所と短所は何か
11. 評価はステークホルダーに評価結果の適切な利用をアシストしたか
12. プログラム評価自体がもたらした結果はどのようなものであったか

概要

1. プログラム評価に関連する一般的課題
2. プログラム評価と研究（評価）デザイン
3. プログラム評価の前提
4. プログラム評価の観点：研究との異同
5. プログラム評価の枠組み

※ 総花的になってしまうことをご容赦下さい

1. プログラム評価に関連する一般的課題

他の関連した評価との関係

- 政策評価
- プログラム（カリキュラム）評価
- コース評価
- 授業評価
- 教員評価
- 学習者評価（評定）
- 自己評価→自己主導型学習

学生からの教員評価

- 教育の顧客からの評価として重要性が高い
- 考慮すべき問題点
  - 目的
  - データ収集
  - 分析

目的

- 報奨制度（Reward system）との関連があるか
- 報奨制度
  - 給与・ボーナス・昇進…
  - 単なる表彰のみ（CVには書けるかも）

➢ 報奨制度と関連しているときは，より高い妥当性が求められる

評価表の作成とデータ収集

例1) ある講師が講義をし，直後に自ら評価表を配布し，自ら集めた

例2) 評価表は非常に細かくて，10ページあった

例3) 医学教育ユニットにおいて，評価表に名前を書いてもらうか否かで議論になった

配布と回収

- 誰が配り，誰が回収したかによって，記載内容は大きく影響を受ける
  - 学習者は脆弱集団（vulnerable population）であることを意識すべき
- 研究・評価のいずれにおいても，誰が配布・回収したかは記載するのがよい

記名・無記名の問題

- 記名
  - フォローアップ可能
  - 回収率低下，回答の歪みが生じる可能性
- 無記名連結可
  - フォローアップ可だが倫理的懸念生じる
- 無記名連結不可
  - 様々な問題は最小限
  - 縦断研究は集団のみ可能に

記名・無記名（Anonymity）

- 記名式アンケートには否定的なコメントを書かない学生は多い（特に教員が集める場合はそうなる）
- 無記名式アンケートには教員側に対して破壊的な否定的コメントをする者も

- 学生による評価目的の十分な理解は前提条件

回収率（Response Rate）

- 高い回収率(70-80%<)は強い結論を導き出すために必要
- 以下の場合学生は記載したくなくなる
  - 評価表が長過ぎる
  - 評価が頻回過ぎる
  - ある授業の総括的評価が終わった後
  - 結果が建設的に利用されていることを学生が知る機会がない（ただでさえ，評価は自分たちには役立たない！）

解析

- 記述統計
- 時系列によるトレンド
  - FD前後の比較などができると有用
- 教員間の比較
  - 変な競争意識はマイナスになることも

評価に利用されるデータ

	主観的	客観的
質的	フォーカスグループ インタビュー	OSCEステーションで問題なく実施された課題の質的分析
量的	量的調査	試験点数のトレンド

Morrison. BMJ. 2003

データの性質

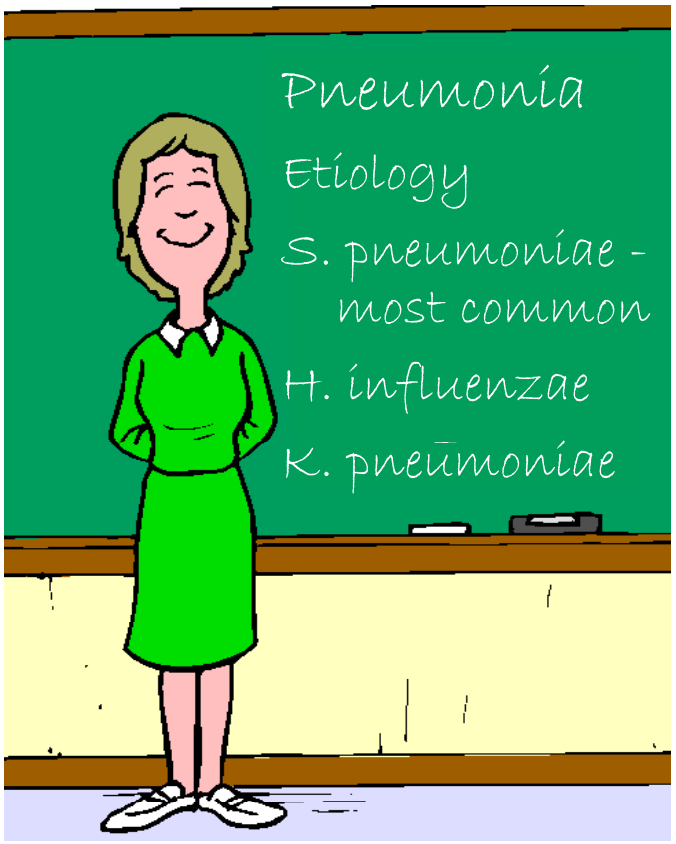
量的データ

- 仮説検証
- 科学的モデルの利用
- 標準化され，客観的で（信頼性のある），再現性を持つ，一般化可能性のある結果の強調
- 学習者のランク付けが可能

質的データ

- より自然な状況で
- 研究者自身を道具として利用可能
- 分厚い記述
- 社会プロセスへの焦点
- 多種のデータ収集方法の利用

質的情報の量化・圧縮



Q: You were satisfied with the lecture.

1) Definitely 2) Yes 3) Somewhat 4) No

1) Her lecture is well organised and easy to understand.

2) Her lecture is understandable, but sometimes too simple.

3) Lecture note is OK, but I hate her artificial smile.

4) Her clothes is too old-fashioned.

東京大学医学教育
国際協力研究センター

2. プログラム評価と研究（評価）デザイン

東京大学医学教育国際協力研究センター

研究デザイン

- 事後テスト単独 vs 事前事後テスト
  - 事前事後テストにより改善度が測定可能
- 対照群の有無
  - 対照群有：準実験
- ランダム割り付けの有無
  - ランダム化有：真の実験

一般的な教育研究デザイン

前実験的 (準実験?)	事後テスト単独	X - - O
	事前事後テスト	O_1 - - X - - O_2
準実験的	非ランダム化比較 事前事後テスト	E O_1 - - X - - O_2 C O_1 - - - - - O_2
真の 実験的	ランダム化比較 事後テスト単独	R E X - - - O_1 C - - - O_1
	ランダム化比較 事前事後テスト	R E O_1 - - X - - O_2 C O_1 - - - - - O_2

X: 教育介入, O: 観察か測定, E: 実験群, C: 対照群, R: ランダム割り付け

前実験的デザイン（広義には準実験に入れることも）の弱点

- 例：生理学の授業に関する効果の確認を試みた
  - O_1 - - X - - O_2：事前事後テストデザイン
  - O_1：55%, O_2：70% (1ヶ月後)
  - 授業は有効だったと結論できるか
- 考慮すべき脅威（Threat）

✓ 履歴効果（学習者に外因が与えた影響）
✓ 成熟効果（時間経過による変化）
✓ 測定効果（測定者や測定法の変化）
✓ テスト効果（1回目のテストが2回目を改善）

準実験デザインの弱点

- 今年から地域の患者を継続的に訪問するプログラムを1年次に開始．今年の2年次（対照群），1年次（介入群）に対して患者の問題に全人的に対応できる能力を口頭試問で評価
- E　O_1 - - X - -　O_2　　非ランダム化比較
  C　O_1 - - - - -　O_2　　事前事後テストデザイン
- 介入群は5段階評価で4.2，対照群は3.1で有意差みられた
- しかし，今回の1年次は女子の率が高く，平均年齢が高かった．おまけに，今年から入試時の面接方法を見直したところだった
- ホーソン効果もみられる可能性あり

各デザインの主な利点・欠点

前実験的	X - - O	状況効果，成熟効果，テスト効果，測定効果は要考慮．介入前後の変化不明
	O_1 - - X - - O_2	介入前後の変化は分かるが他の効果は同様
準実験的	E　O_1 - - X - - O_2 C　O_1 - - - - -　O_2	状況効果，成熟効果，テスト効果，測定効果は排除可．選択バイアスは要考慮
真の実験的	R　E　X　- - - O_1 C　- - - O_1	選択バイアスは排除可．介入前後の変化不明．ランダム化が上手くいったか不明
	R　E　O_1 - - X - - O_2 C　O_1 - - - - -　O_2	ランダム化が上手くいったかどうか事前評価可能．対照群は教育介入受けない

更に難しい点

- 対照群を設けること(教育しない群ができる),
  ランダム割り付けすることは倫理的にOK？
- 対照群は人口動態学的要因，心理学的要因
  に関して実験群と同等か
- 全国展開するプログラムには比較群無し

→ 真の実験デザインを選択すること自体が困難

倫理的懸念を避けるデザイン

- ランダム化比較事前事後テスト

R E O_1 - - X - - O_2
R C O_1 - - - - - O_2

※対照群の学生は教育介入を受けられない

- 交代介入による比較事前事後テストデザイン

R E O_1 - - X - - O_2 - - - - - O_3
R C O_1 - - - - - O_2 - - X - - O_3

広義の準実験デザインの工夫例

- 複数指標による1群事前事後テストデザイン

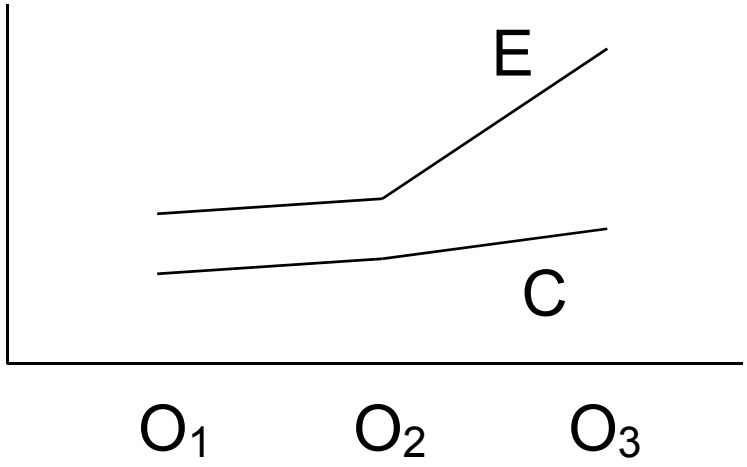
O_{1A} O_{1B} X O_{2A} O_{2B}

  - O_{1A}とO_{1B}に差がなく，O_{2A}とO_{2B}にも差がない
  - O_{1A}とO_{1B}の平均とO_{2A}とO_{2B}の平均との間には有意差がある
  - Xがこの差をもたらしたと推測可能
  - 履歴効果，成熟効果，測定効果，テスト効果は，いずれもゼロではないが，単なる事前事後テストデザインよりははるかによい

準実験デザインの工夫例

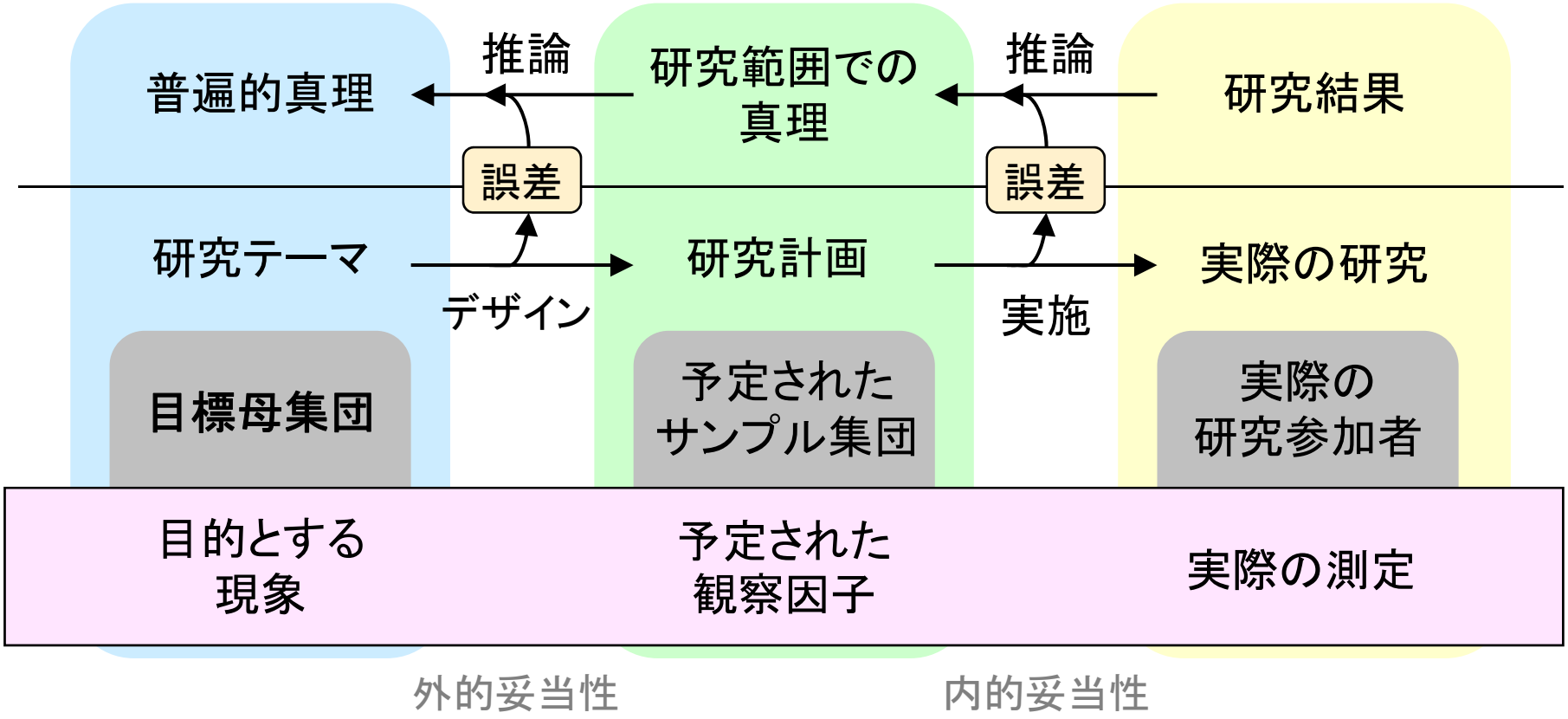
- 複数回の事前テストによる不等価統制群事前事後テストデザイン

E	O_1	O_2	X	O_3
C	O_1	O_2		O_3



- 介入群において，介入後の傾きの変化が大きい

デザイン，実施，推論の対応



(木原監訳．医学的研究のデザイン．p35)

どこまで一般化可能？

- 各国・各大学のカリキュラムに依拠してしまう研究・評価
- 多施設研究，多国間研究なら解決？
- 医学教育分野では，研究対象集団を便宜的に決め，読み手の解釈を許す形で記載されたものが多い印象

実験的方法論を用いるには

- 評価仮説（せいぜい数個）が必要
- 介入の対象・性質からデザインを上手く構築する必要がある

3. プログラム評価の前提

プログラム評価のあり方

- 単に良い・悪いと一元的に決められる評価など，きっとほとんどない
- 現状の改善，失敗からの学習，次回に活かす・・・
- 誰が，何を，誰のために，何のために，どのように，どのような枠組みにて，どの程度のコストで，どの程度まで，評価すればよいのか？

Evaluation vs Assessment

- 評価：Assessment
  - 学生の達成度，能力の査定，評定，評価
  - プログラムのニーズ
- 評価：Evaluation
  - 教員や授業
  - プログラムやカリキュラム
  - 学部や大学全体
  - 政策レベル

プログラム評価の定義

- Scriven (1967)
  - 方法論的な活動により，実績データを目標値と照らし合わせること
- Scriven(1991)
  - 明確化および正当化された基準により，物事（つまりプログラム）の価値を判断すること

➢ 以前よりはプログラムのvalueを重視する定義が増えている印象

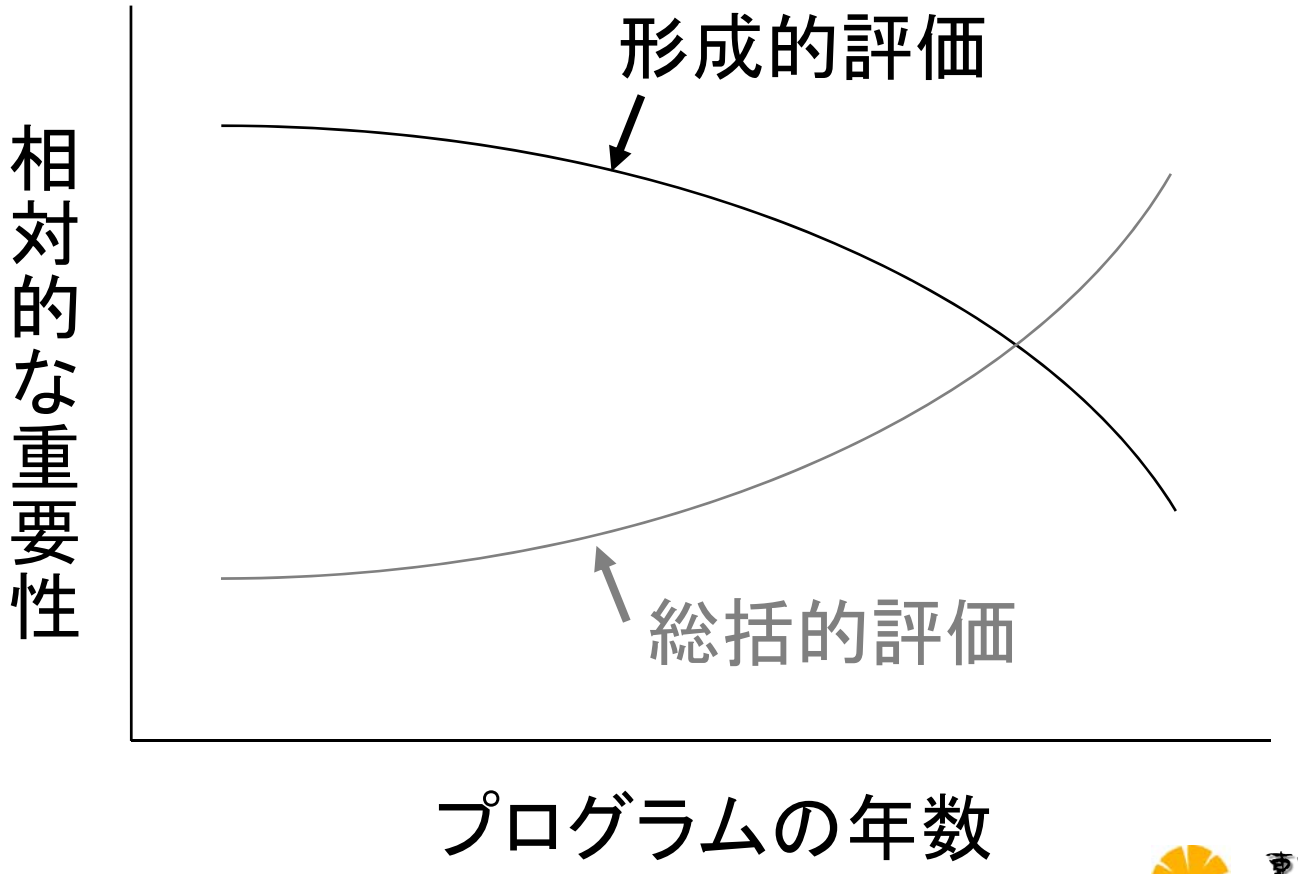
何を評価するか

- プロセス（process）
  - 工程（operation）
  - 機能（function）
- アウトカム・結果（outcome）
  - 効果（effectiveness）
  - アウトプット（output）

プログラム評価の目的

- 形成的評価（formative evaluation）
  - 目的：プログラムの発展と改善
- 総括的評価（summative evaluation）
  - 目的：アカウンタビリティ
- 価値判断や意思決定のための評価
- 評価を通じた知識の習得や付加
- 宣伝活動のための評価

プログラムの寿命と評価目的の変化



内部と外部の目

- 外部からの評価・分析
  - ✓ 内部の政治から中立
  - × 一部の情報へはアクセスが困難
- 内部からの評価・分析
  - ✓ 情報は豊富
  - ✓ 評価が学習につながる
  - × 重要な側面が死角になる／隠蔽されることがある

4. プログラム評価の観点
～研究との異同～

よくみられる事例

（例）クリニカル・クラークシップに対して，
指導医・後期研修医・初期研修医・
医学生からなるチーム管理手法を
新たに導入した．この手法が有用で
あることを学会発表する

→　学会発表≒研究（宣伝？）

科学的側面と実用的側面

- Donald Campbell (1969)
  - 政策やプログラムに関する決定は，社会状況を改善するための方法を検証する継続的な社会実験から引き出されるべき
- Lee Cronbach (1982)
  - 評価も科学的研究と同様の科学的手続はとるが，評価の目的は科学的研究から明確に区別されるべき．利害関係者のニーズ充足を志向すべき．

そもそも研究とは？

- 今までに誰にも知られていないことを解明していく
- 広く一般化可能な“知”を生み出す

- 医学教育に関する学会発表の一部
  - 自分による，自分のプログラムに対する，自分のための評価？

探求の焦点

- 研究
  - 過去の理論に基づいて構築
  - 新たな仮説の創出と検証
  - 一般化可能性を意識
- 評価
  - 次の政策やアクションにつなげる
  - プログラムの価値を同定

結果の一般性

- 研究：より一般性を求める
  - 研究デザイン
  - 母集団
  - 対照集団の選び方
- 評価：より結論を求める
  - 効率
  - バランス
  - ステークホルダーの選び方

重要性の基準

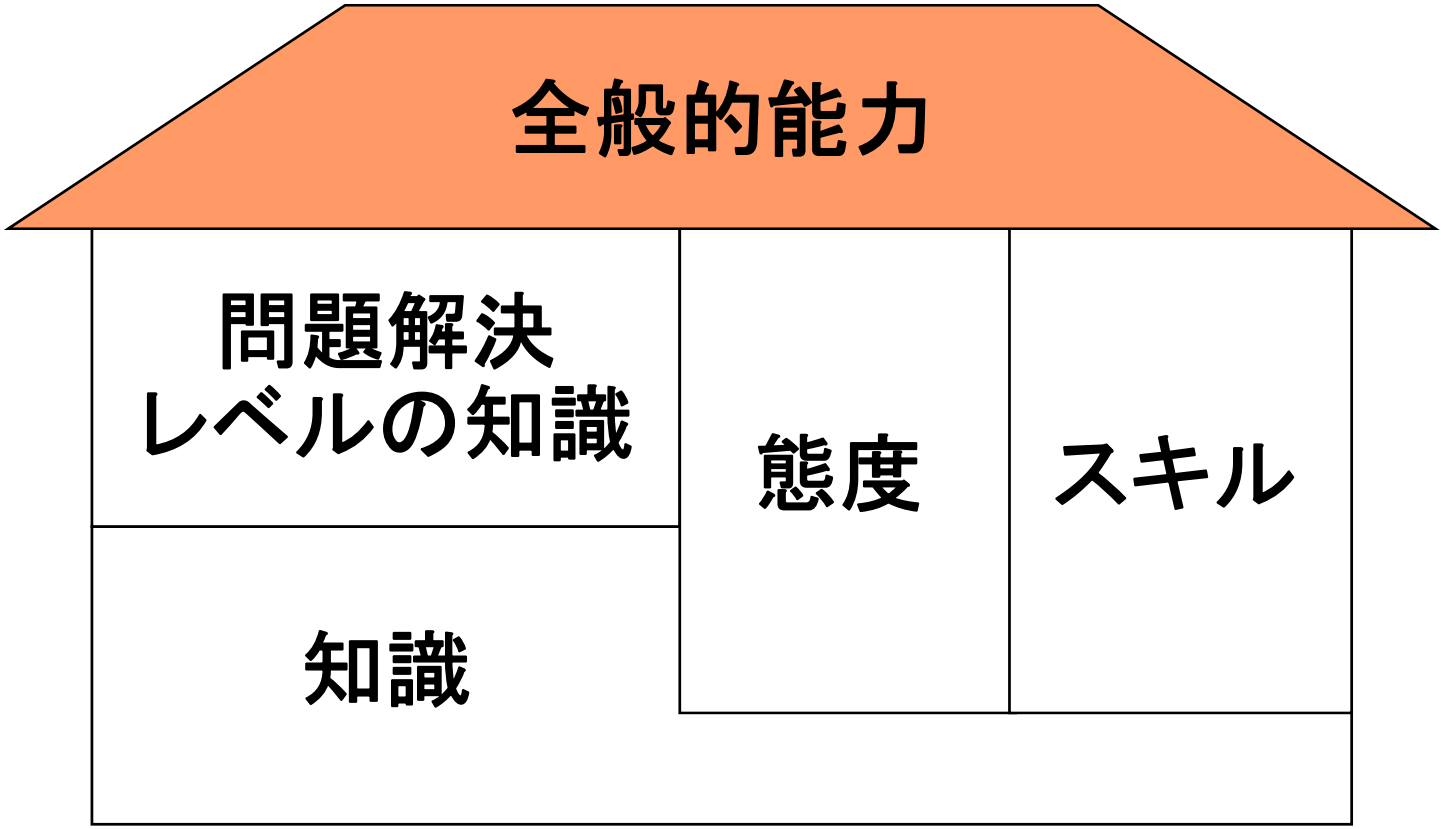
- 研究
  - 真実（truth）の追究
- 評価
  - 価値（value）の追求

研究と評価の違い

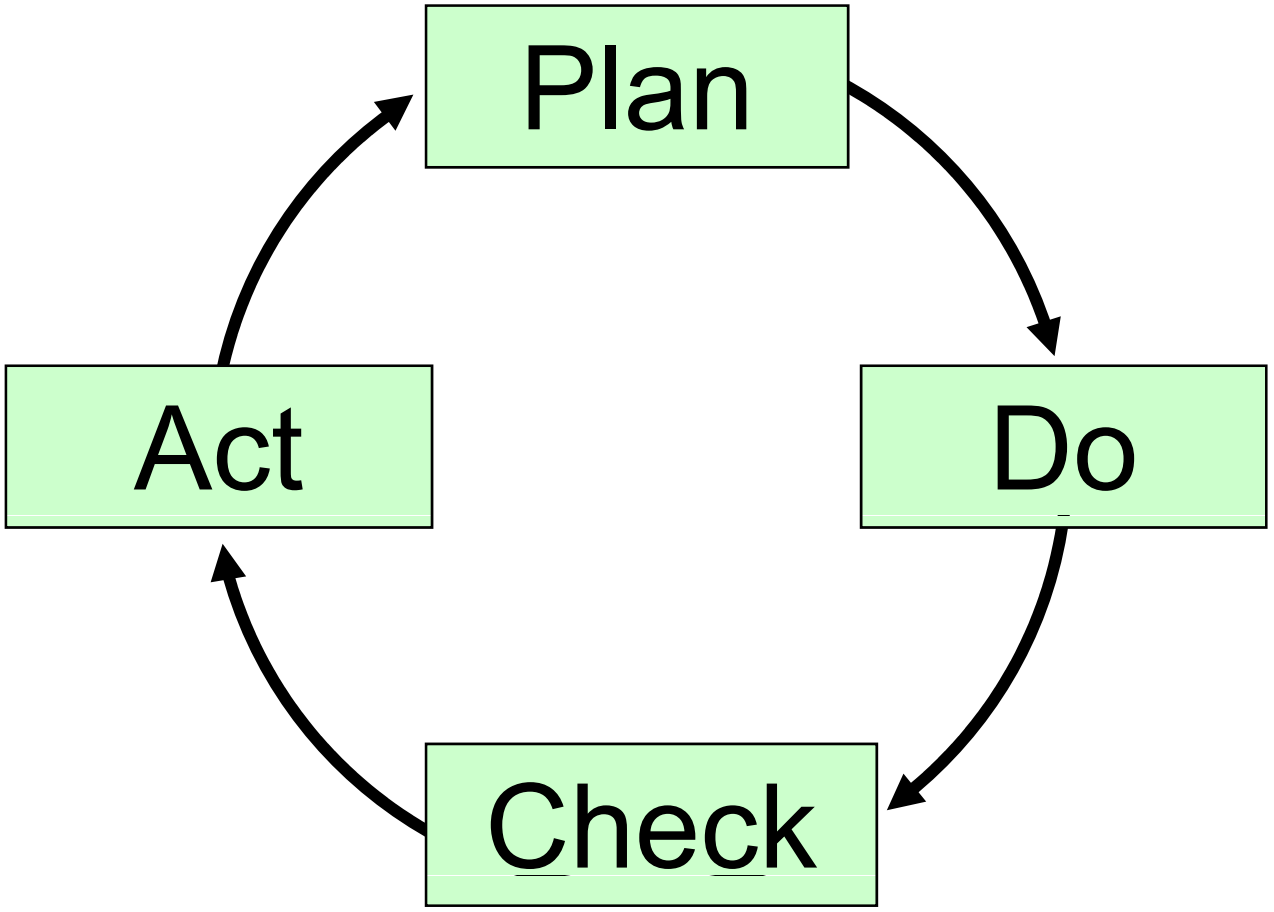
	研究	評価
探求の焦点	考察	決定
結果の一般性	高	低
重要性の基準	真実	価値

5. プログラム評価の枠組み

全般的能力と個別目標の関係



評価（マネジメント）サイクル



カリキュラム開発モデル

6段階アプローチ（Kernら．1998）

1. 問題の明確化と一般的ニーズ評価

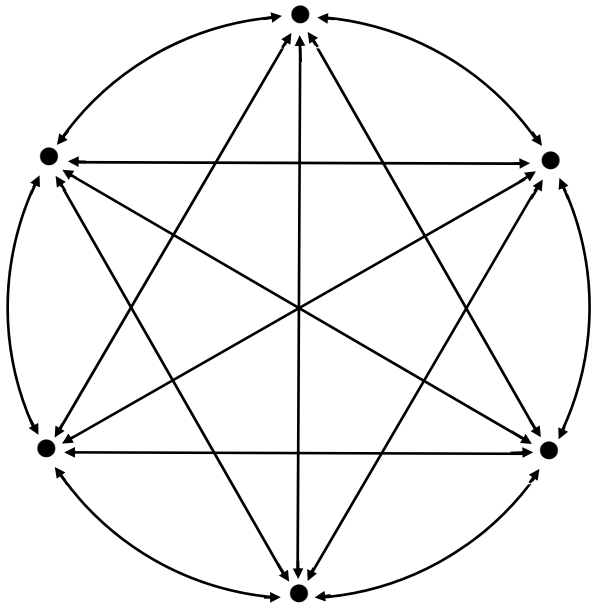
2. 対象学習者のニーズ評価

3. 一般目標と個別目標

4. 教育方略

5. カリキュラムの実施

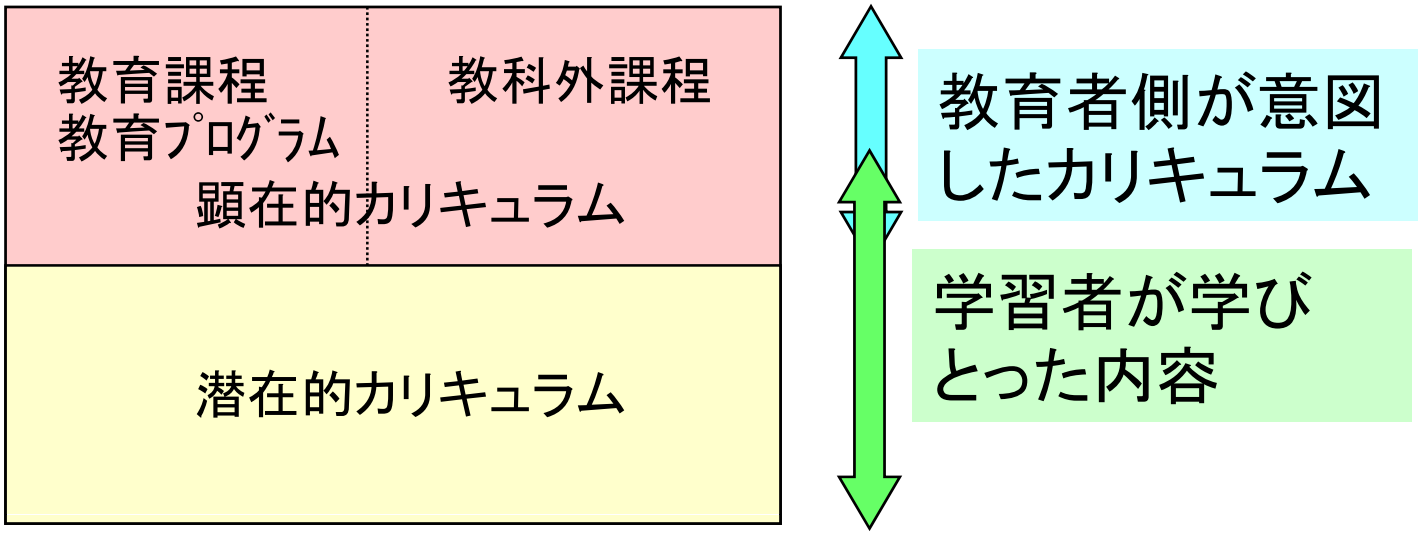
6. 評価とフィードバック



東京大学医学教育国際協力研究センター

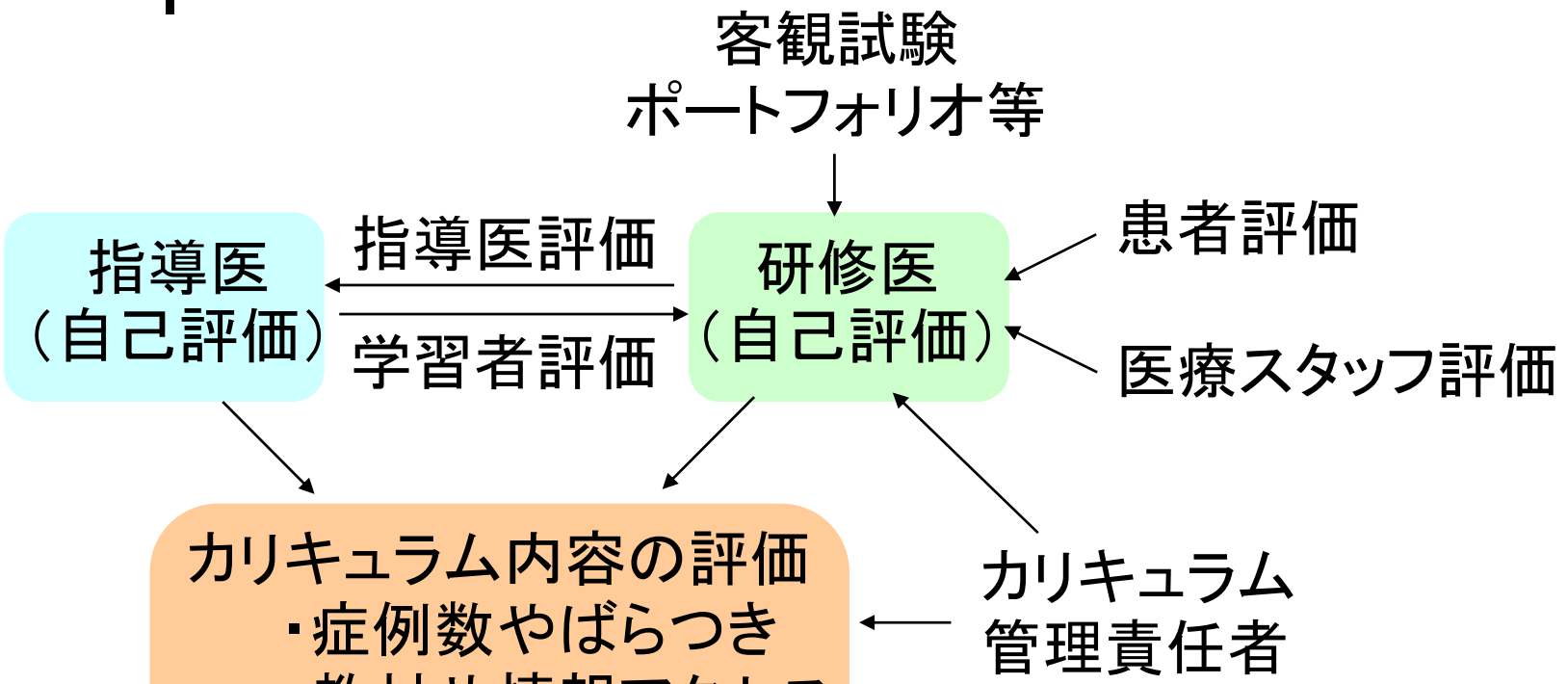
カリキュラムの捉え方

- カリキュラム：学習者が学びとった内容



  - カリキュラムとは，スケジュールや教育方法だけでなく，患者とのやり取り，耳学問，自己学習内容など全て

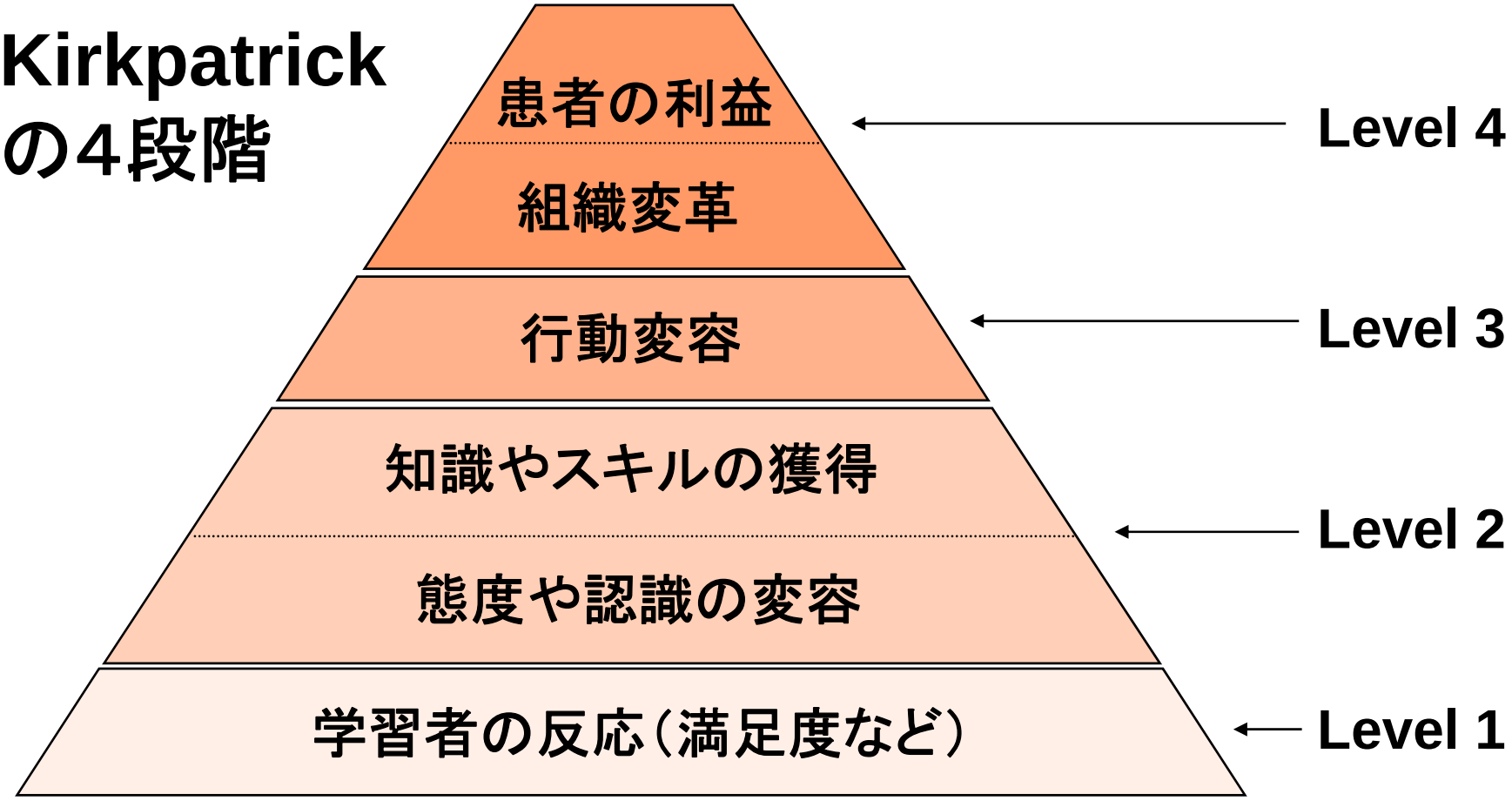
臨床研修評価の要素



これらの内容は，プログラム内部の項目のみを捉えていることに注意．例えば，育った研修医が社会にどのような影響を及ぼすかなどは含まず．

評価の焦点のレベル

Kirkpatrick
の４段階



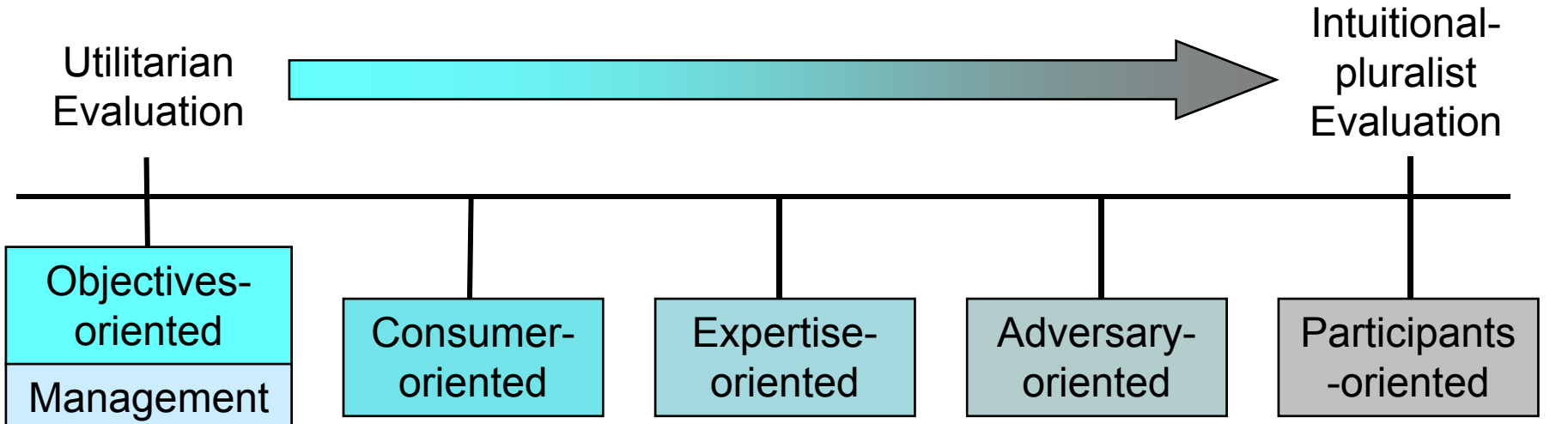
カリキュラム評価論

- 行動目標的アプローチ（Tyler RW）
- ゴール・フリー評価（Scriven M）
- 羅生門モデル（Atkin JM）
- 教育批評（Eisner EW）

マネジメント志向型評価
～CIPP評価モデル(stufflebeam)～

- Context：計画を評価．ニーズ調査
- Input：システムを評価．リソース調査
- Process：実施段階を評価．開始後の問題点や改善点を調査
- Product：アウトカムを評価．当初計画と比較し見直しを図る

評価アプローチの観点



Objectives-oriented / Management-oriented

- Tylerian Evaluation Approach
- Logical method

- Stufflebeam's CIPP Evaluation Model

Consumer-oriented

- Scriven's model (goal free ev)

Expertise-oriented

- Formal/ informal Professional Review
- Ad Hoc Panel/ individual Review
- Educational Connoisseurship and Criticism

Adversary-oriented

- Owen's model
- Judicial ev

Participants-oriented

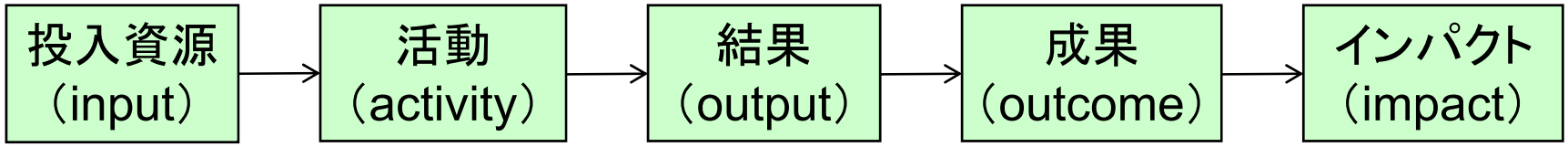
- Stake's Countenance Model
- Transactional ev
- Illuminative ev
- Democratic ev
- Naturalistic ev
- Reponsive ev

OECD開発援助委員会 評価5項目（ODA評価）

妥当性 (relevance)	有効性 (effectiveness)	効率性 (efficiency)	インパクト (impact)	自立発展性 (sustainability)
プロジェクトの計画内容は受益者のニーズと合致しており必要性が高いか、相手国の開発政策および日本の援助政策と整合性があり高い優先度が認められるか、対象分野・セクターの問題や課題の解決策として適切か。	プロジェクトで計画した効果（プロジェクト目標）は達成されているか、それはプロジェクトの活動の結果もたらされたものか。	アウトプットもしくはプロジェクト目標について、より低いコストで達成する代替手段はなかったか、あるいは同じコストでより高い達成度を実現することはできなかったか、投入はタイミングよく実施されたか。	プロジェクトで計画した長期的・間接的な効果（上位目標）は達成されているか、予期していなかった社会経済的な正・負のインパクト（波及効果）はあるか。	プロジェクトが目指していた効果（プロジェクト目標、上位目標）は協力終了後も持続するかについて、技術・組織・財務などの視点から。

この方法論から学ぶべきこと

- management-orientedな手法の一つ
- ロジックモデル（PCM：project cycle management）を用いており，活動・結果・成果・インパクトの関係が明確



- カリキュラム実施前に詳細計画を煮詰め，その計画に基本的に沿って実施することが必要
- 出資者，相手国政府など外交的な配慮も必要．多方面のステークホルダーの意見が一致することが最重要．評価者は各ステークホルダーとのコミュニケーションを十分に図ることで，バランスを取った結論を出す

特に教育プロジェクトで重視される点

- インパクト（impact）
  - 正だけでなく，負のインパクトも評価していくことで，ロジック・前提条件などに立ち返ってフィードバックできる
- 持続発展性（sustainability）
  - 技術・組織・財務の3側面から持続発展の可能性を探ることで，変革が成功したかどうかを確認

まとめ

- データ収集からプログラム評価の枠組に至るまでの概要
- 実験的手法による客観性の重視だけでなく，ステークホルダーのニーズを満たすことができるという視点も重要
- 研究と評価は異なるが，研究に関連した経験や知識は非常に重要