What Does It Mean to Do Rigorous Research in Education? A discussion of Theory and Methods

Alan H. Schoenfeld
University of California
Berkeley, CA, USA
Alans@Berkeley.edu

Main Themes of This Talk

1. Purposes
   - Goals of research in education
   - The relationship between theory and practice
2. The question of values

   How values shape what we do, and why it's important
3. Research Methods:
   - A framework for conceptualizing the process
   - Principles for conducting high quality research

1. The Purposes of Research in Education

Pure (Basic Science):

To understand the nature of thinking, teaching, and learning;

Applied (Engineering):

To use such understandings to improve instruction.

Examples of pure research:

Constructivism – Piaget

The magic number 7±2 – Miller

Arithmetic bugs – Brown and Burton

Behaviorism – Skinner

Expertise - deGroot

On the relationship between research and practice:

For many years, the idea was to do “basic research” and then apply it.

But classrooms are more complex than laboratory experiments! Often, moving from theory to applications doesn’t work.

There’s another way, doing basic work in context. Stokes called this “Pasteur’s Quadrant.”

Research is inspired by:

Quest for Fundamental Understanding?	Consideration of Use? No	Consideration of Use? Yes
Yes	Pure Basic Research (Bohr)	Use-Inspired Basic Research (Pasteur)
No		Pure Applied Research (Edison)

Today we conduct “design experiments,” working at the edge of theory and practice.

In a design experiment, you have a theory of what will work, and why; empirical research refines both the theory and the intervention.

2. The question of values

We are often asked questions like,

- Are large classes as good as small classes?
- Is curriculum A better than Curriculum B?
- What is the right way to teach X?

These questions seem to make sense, but they are unanswerable. The answers all depend on values - on what you consider to be the most important outcomes.

Consider the question,
"Are large math classes as good as small classes?"

How will you judge the outcomes? Will you look at:

- Scores on a multiple-choice final exam?
- Skills at modeling and problem solving?
- How many students continue studying mathematics?

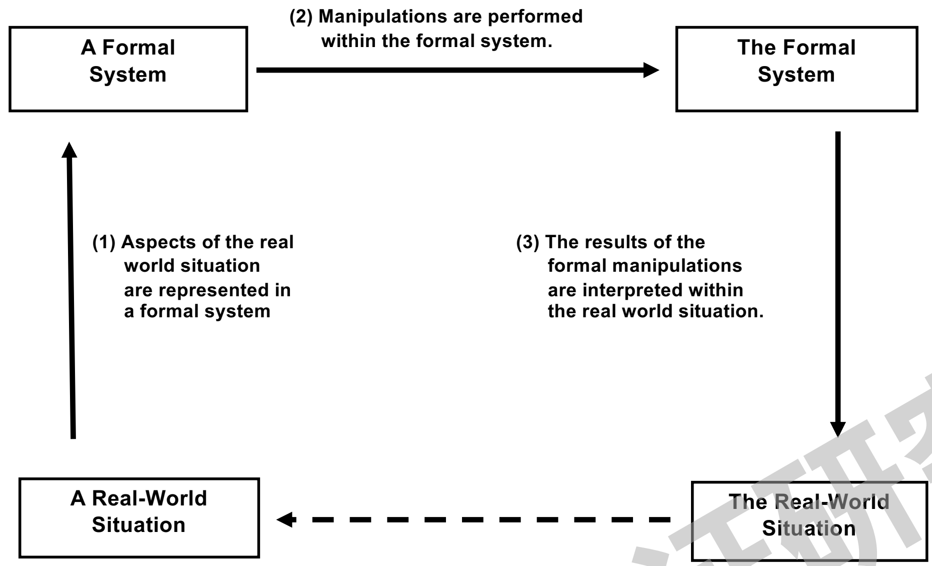
The answer to "what is best" depends on what you value – and how you measure it.

For example, countries around the world have very different rankings on TIMSS and PISA.

That's because TIMSS and PISA test different things. They reflect different mathematical values.

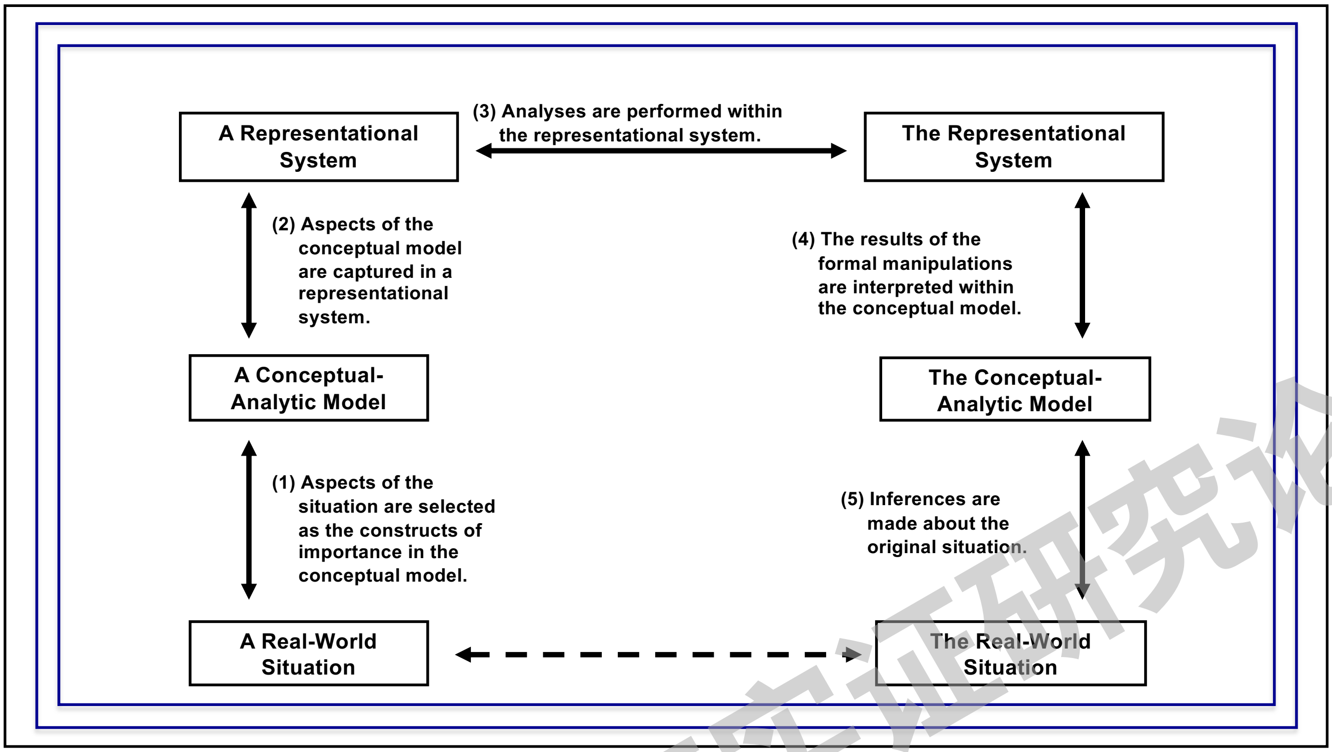
3a. A Framework for Conceptualizing the Research Process

The next slide shows how most people (including mathematicians and scientists) tend think of the representation and modeling process.



The reality is much more complex.

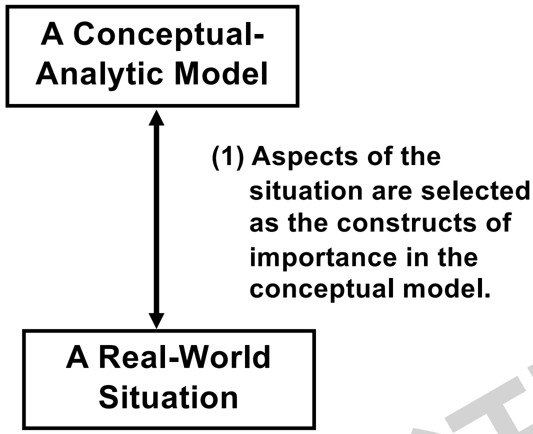
This is the process as a whole:



There are challenges and complexities at every step.

I will discuss the transitions one at a time.

The First Transition

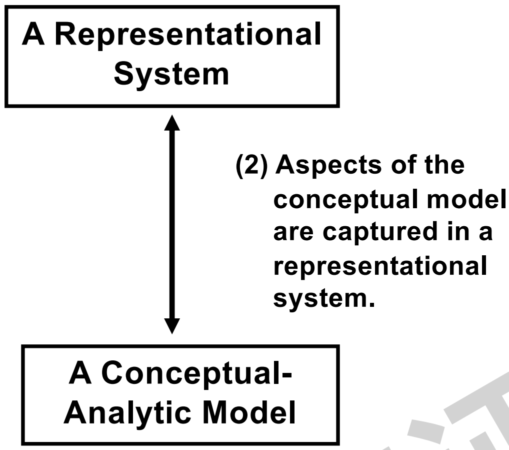


What we *select* in order to frame and conceptualize the "real world" situation in theoretical terms is fundamentally important.

For example:

- Who do the "subjects" represent?
  (Think of medical studies, of men only)
- What "counts" as understanding?
- What classroom processes are of interest?

The Second Transition



For example,

How will you decide whether students have learned elementary statistics?

Consider these two problems.

Problem 1:
Compute the mean, median, and standard deviation of these two distributions, and represent them graphically:

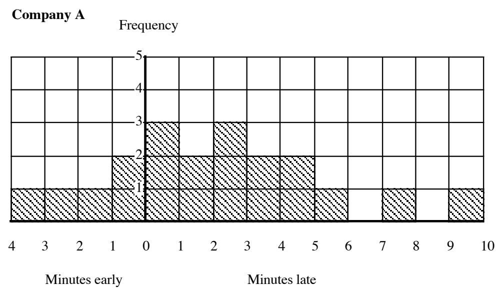
a. -3.5, .75, 1.5, 4.5, -.75, -2.5, 4.75, 2.75, .5, -1.5, 2.25, 9.25, 3.5, 1.25, -.5, 2.5, .5, 7.25, 5.5, 3;

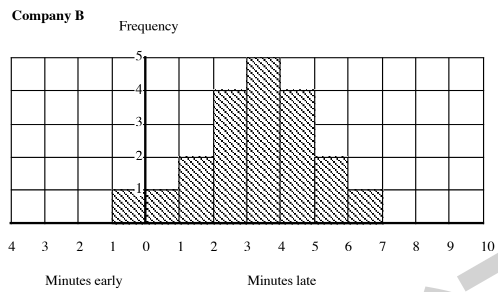
b. 3.75, 4.5, 3, 5, 2.25, 1.25, .75, 3, -.5, 1.5, 3.5, 6, 4.5, 5.5, 2.5, 4.25, 2.75, 3.75, 4.75

Data:

	Distribution A	Distribution B
Mean	2.3	3.14
Median	1.53	3.15
Range	12.45	6.3
SD	3.11	1.4

Graphs:





This is a complete answer.

Problem 2:

You work for a business that has been using two taxicab companies, Company A and Company B.

Your boss gives you a list of (early and late) "Arrival times" for taxicabs from both companies over the past month.

Your job is to analyze those data using charts, diagrams, graphs, or whatever seems best. You are to:

i. make the best argument that you can in favor of Company A;

ii. make the best argument that you can in favor of Company B;

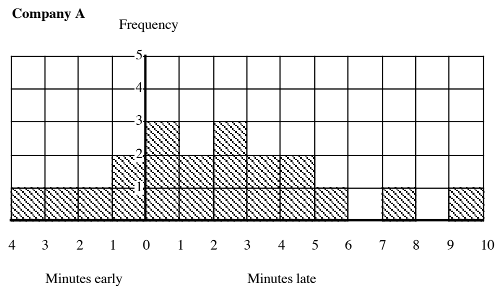
iii. write a memorandum to your boss that makes a reasoned case for choosing one company or the other, using the relevant mathematical tools at your disposal.

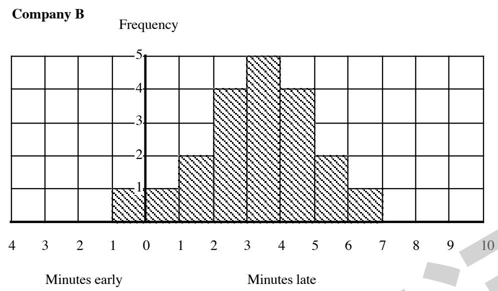
Company A		Company B	
3 mins 30 secs	Early	3 mins 45 secs	Late
45 secs	Late	4 mins 30 secs	Late
1 min 30 secs	Late	3 mins	Late
4 mins 30 secs	Late	5 mins	Late
45 secs	Early	2 mins 15 secs	Late
2 mins 30 secs	Early	2 mins 30 secs	Late
4 mins 45 secs	Late	1 min 15 secs	Late
2 mins 45 secs	Late	45 secs	Late
30 secs	Late	3 mins	Late
1 minute 30 secs	Early	30 secs	Early
2 mins 15 secs	Late	1 min 30 secs	Late
9 mins 15 secs	Late	3 mins 30 secs	Late
3 mins 30 secs	Late	6 mins	Late
1 min 15 secs	Late	4 mins 30 secs	Late
30 secs	Early	5 mins 30 secs	Late
2 mins 30 secs	Late	2 mins 30 secs	Late
30 secs	Late	4 mins 15 secs	Late
7 mins 15 secs	Late	2 mins 45 secs	Late
5 mins 30 secs	Late	3 mins 45 secs	Late
3 mins	Late	4 mins 45 secs	Late

Same Data:

	Distribution A	Distribution B
Mean	2.3	3.14
Median	1.53	3.15
Range	12.45	6.3
SD	3.11	1.4

Same Graphs:





But here's the right kind of answer:

"Company A's cabs are earlier on average than Company B's, but they are less consistent in their arrival times. It's better to order a cab from Company B - but order it for 5 minutes early, so it arrives when you need it."

Both questions test statistical competency, but in *very* different ways. Students can do well on the first (skills) and not on the second (understanding and use).

Because of differences like these, tests can misrepresent what kids know and can do.

E.g., two tests of more than 5000 7th graders:

Skills Test (SAT-9)		Broad Test (Balanced Assessment): Pass	Broad Test (Balanced Assessment): Fail
	Pass	32%	28%
	Fail	2%	38%

The Third Transition

A Representational System ⟷ **The Representational System**

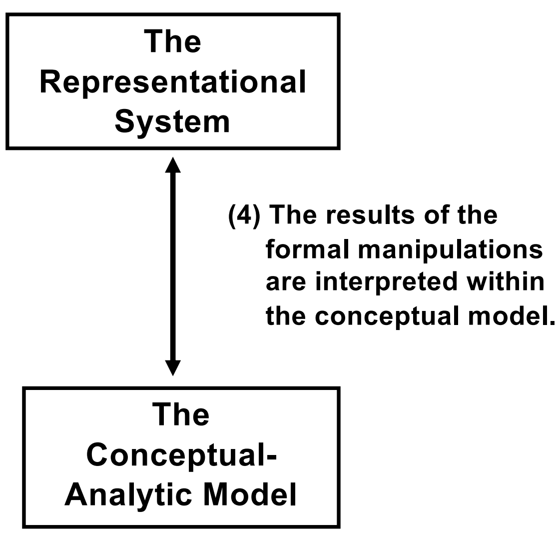
(3) Analyses are performed within the representational system.

The nature of the analyses (and their suitability and interpretation) may or may not be appropriate.

For example:

- Do the conditions for analysis match the assumptions in the model?
- Are the phenomena captured in robust and meaningful ways?
- Are the analytic methods stable (e.g., do they have good inter-rater reliability)?

The Fourth Transition

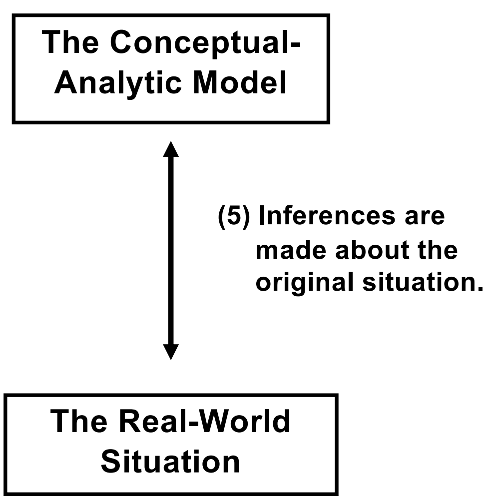


"Mapping back" to the conceptual model may or may not make sense.

For example:

- Sampling error - Dewey beats Truman!
- Construct validity - e.g., what is "IQ" or "Power relationships, or "self concept?
- Econometric analyses - do proxies for "the ratio of instructional staff to students" really make sense?

The Fifth Transition



The application of the process to the "real world situation" can be problematic.

For example:

- Do psychometrically defined entities such as "verbal or spatial ability" make sense?
- Can ideas such as "sociomathematical norms, or "knowledge," "goals," "beliefs," be defined and used in meaningful ways?
- What's the unit of analysis? A reporter at *Science* misunderstands completely:

A superintendent did a study where he compared three "experimental" schools with three "control" schools, and got no differences on average.

One of the schools worked with the program and got great results, one was so-so, and one resisted the program and got lousy results.

What's the moral of the story?

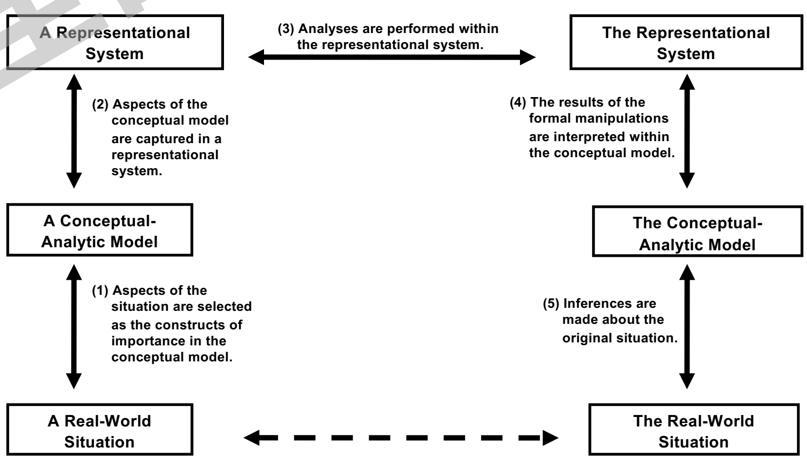
The reporter said, “there was no difference between treatments.”

But as I see it:

There are three conditions: supportive, neutral, and negative.

If the conditions are supportive, use the new curriculum. If they’re neutral, provide training and encouragement. If they’re negative, stay with the old treatment.

In sum, every aspect of this diagram:



must be considered with great care, both in designing our own research and in making sense of the research done by others.

3b. Principles for Conducting Research

In order to conduct high quality research that can have useful applications,

1. Both theory and methods should be held accountable to data;
2. Research should be judged along at least three dimensions:
   - trustworthiness
   - generality
   - importance.

(None of these guarantees any of the others.)

Accountability to data.

As a field, we are much too casual in our use of data, especially qualitative data. We tend to formulate our ideas about what is important, and then offer segments of data that substantiate our ideas ("confirmation bias").

We must gather data that allows for the possibility of proving ourselves wrong, and we must find ways to deal comprehensively with all the data we gather.

There is also the matter of how we hold ourselves accountable. My own personal standard for qualitative data is to be able to model people's actions and decision-making. The reason is that such modeling guards against ad hoc justifications, and forces me to try to explain everything, or explain why I can't.

From my personal perspective, model-building as a form of theory testing is essential.

This has serious implications for how we report data, if we want our work to build and be cumulative.

It is essential for us to provide much more data, and much more detail on our analytic methods, than has been the tradition. This should allow for (a) alternative interpretations, and (b) replication.

Dimensions for Judging research:

- trustworthiness
- generality
- importance.

Aspects of trustworthiness include:

- Descriptive and explanatory power
- Prediction and falsification
- Rigor and specificity
- Replicability
- Triangulation

In fact, our whole "tour" of the diagram was devoted to issues of trustworthiness.

Generality

Beware of claims about generality – there are different types:

The *claimed generality* of a body of research is the set of circumstances in which the author of that work claims that the findings of the research apply.

The *implied generality* of the work is the set of circumstances in which the authors of that work appear to suggest that the findings of the research apply.

The *potential generality* of the work is the set of circumstances in which the results of the research might reasonably be expected to apply.

The *warranted generality* of the work is the set of circumstances for which the authors have provided trustworthy evidence that the findings do apply.

Importance

(I shouldn’t have to say much here!)

Studies should work toward the purposes discussed at the beginning of this talk:

Pure (Basic Science) – To understand the nature of mathematical thinking, teaching, and learning; and

Applied (Engineering) – To use such understandings to improve mathematics instruction.

Finally, is there “one right way” to do high quality educational research?

Of course not!

There are various forms and methods of rigorous research.

These may correspond to:

- Large-scale statistical studies (properly conducted and interpreted)
- Individual findings correlated to large studies
- Replications over time
- Detailed analytic models tied to theory
- Predictions tied to theory and models

Large-scale statistical studies (properly conducted and interpreted)

For example,

TIMSS and PISA test results (when you know what the tests test).

Individual findings correlated to large studies, for example:

My studies of student beliefs in one classroom, backed up by questionnaires given to 240 students, which demonstrate they are typical, or triangulated with results on the US National Assessment of Educational Progress.

Replications over time:

- Thousands of students with the same "read/explore" pattern as in the graphs I showed in my talk on problem solving. (More than 50% of all students, over many years.)
- It's easy to replicate Miller's "magic number 7±2". Similarly for perception of illusions, vertical translations, etc.

Detailed analytic models tied to theory:

- Doing this kind of work is the same as in any empirical science: you construct theories, and you build models to test them, comparing the output of the models to the phenomena being modeled.

Predictions tied to theory and models:

- Brown and Burton predicted students' *incorrect* answers to subtraction problems, before the students worked the problems!

Our responsibility is to use appropriate methods – often multiple methods – for the issues we are investigating…

… and then make careful claims that to not go beyond what the results indicate.

(Though we can always suggest what we think will be true.)

In Summary…

- Education is fundamentally concerned with theories of thinking, teaching, and learning.
- Education is fundamentally concerned with the improvement of educational practice.
- Like any empirical discipline, it profits from the exchange between theory and practice.
- As an empirical discipline, it adheres to the standards for the empirical sciences. We must do rigorous work, with well established norms and standards.

謝謝