INSTRUMENTATION AND INNOVATION IN DESIGN EXPERIMENTS: TAKING THE TURN TOWARDS EFFICIENCY

Daniel L. Schwartz, Jammie Chang, Lee Martin

Stanford University

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grants No. BCS-0214549, REC-0196238, & SLC-0354453. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Corresponding Author:

Daniel L. Schwartz
School of Education
485 Lasuen Mall
Stanford, CA 94305
Email: Daniel.Schwartz@stanford.edu
Phone: (650) 736-1514

A design experiment is a form of interventionist research that creates and evaluates novel conditions for learning. The desired outcomes include new possibilities for educational practice and new insights on the process of learning. Design experiments differ from most educational research, because they do not study what exists; they study what could be.

Methods for determining "what could be" are underdeveloped. A science of novel intervention needs both practical methods for doing productive research and logical methods for evaluating research. Some authors propose that intervention research should adopt product design methodologies that include iterative cycles of mid-stream modification, retrospective sense making, case studies, and human sensibilities (e.g., Collins et al., 2004). These methods are good for making products, but they are not ideal for producing generalizable causal knowledge. Others propose that intervention research should resemble classic experiments that emphasize random assignment, dispassionate analysis, and hypothesis testing (e.g., Shavelson, Phillips, Towne, & Feuer, 2003). These approaches are good for creating generalizable knowledge, but they are not practical for early stages of innovation.

These methodological camps are often set apart as opposites, and at best, some researchers imagine that they can bridge the gap by jumping from observational design methodologies to large-scale clinical trials. As we discuss below, this jump is rarely optimal. We find it more productive to re-characterize the methods in a larger learning space that arose from an analysis of the ways people generalize learning to new contexts (Schwartz, Bransford, & Sears, 2005). The learning space has two axes (Figure 1a). The

horizontal axis represents processes and outcomes associated with efficiency. The vertical axis represents innovation. Below, we say much more about these dimensions of learning. For now, Figure 1b shows the learning space adapted to science.

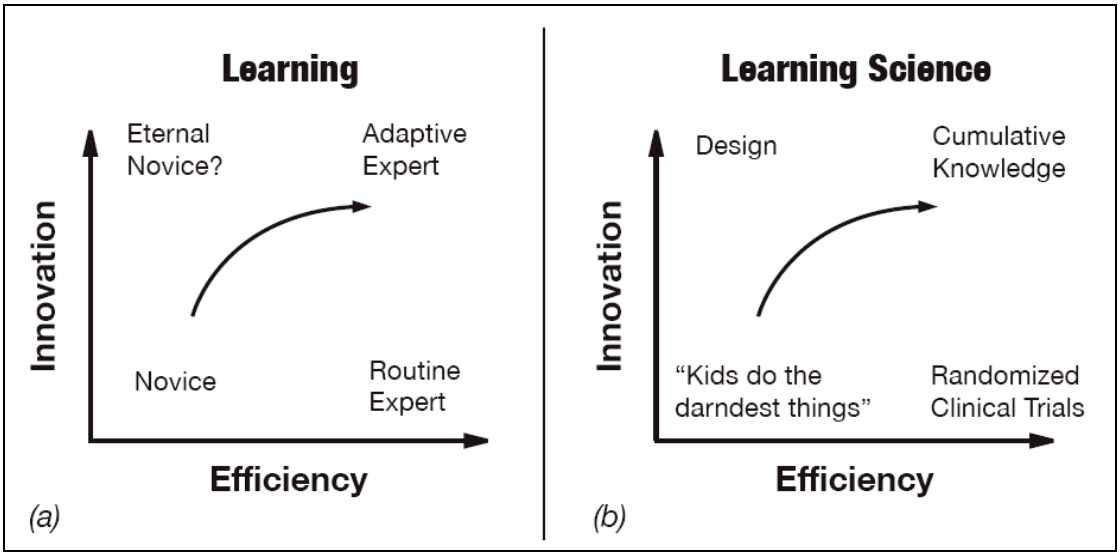

Figure 1. Trajectories of knowledge growth. (Schwartz, Bransford, & Sears, 2005).

Design methodologies are high on the innovation dimension, but low on efficiency: their goal is often discovery and the creation of novel practices, but they are poor at developing efficient tests and descriptions of causal hypotheses about learning. Clinical trials are high on the efficiency dimension, but low on innovation: they test hypotheses about which intervention causes superior learning, but they require sample sizes that are too costly for vetting every innovative idea. Good science needs a balance of innovation and efficiency. In the upper-right corner, we position the ultimate goal of science – the cumulative growth of knowledge – which requires innovation and efficiency.

Given the space of innovation and efficiency, the question is not which

methodology is right. The question is how to promote movement through the space. The arrow in the figure reflects our proposal that an optimal trajectory first accelerates along the innovation dimension and then turns towards efficiency. Science would hardly advance if people merely tested and refined pre-existing hypotheses. However, innovation-oriented research needs to include provisions for a turn to efficiency, lest the field get mired in isolated innovations without prospect of accumulation.

The methods we use in our research attempt to facilitate this movement through the design of research instruments. Though we design technologies, classroom structures, lesson plans and the like, we think of these primarily in terms of instrumentation. Most of our lab discussions involve creating, evaluating, and calibrating instruments. We are not alone – instrumentation is central to all science. We asked a physicist how much of his research involves conducting experiments. He responded that 95% of his time is spent on instrumentation (Jose Mestre, 2004, *personal communication*), by which he meant both the design and calibration of apparatus to precipitate effects and the methods to measure those effects. In our experience, the ability that differentiates novices from domain experts in science is the ability to determine the right instruments for the task.

Designing instruments provides a bridge between innovative design and efficient experimentation. In our work, we expend great effort testing the innovative value of our instruments through *assessment experiments*. In the primary example below, we argue that current instruments to promote and measure learning often miss people's *preparation for future learning* (Bransford & Schwartz, 1999). To make this argument, we innovated both the apparatus that prepared students to learn and the measures that evaluated this

preparation. This work did not involve isolating or proving causality. Rather, we wanted to show that our instruments reveal something that others do not. At the same time, these instruments, which generated and measured a phenomenon, equip us for subsequent efficiency research that can address issues of causality and generality.

We divide the chapter into four sections. In the first section, we develop the case for considering learning and science in terms of innovation and efficiency. In the second section, we argue that instrumentation can encourage both innovation and efficiency in learning and science. In the third section, we show how instrumentation research can support both student learning and science through a trajectory of innovation that turns towards efficiency. In the final section, we turn to the problem of knowledge warrants in design experiments. Efficient science has a number of formal techniques for warranting claims of knowledge progress (e.g., p-values, control conditions). Can design experiments be evaluated by warrants of progress in innovation before these innovations are mature enough to stand the tests of efficient science? We distinguish innovations in knowledge from innovations in practice, and our solution highlights the belief that criteria of efficient science depend on predicting future regularities, whereas criteria of innovation depend on reconciling past irregularities.

THE CASE FOR THE EFFICIENCY AND INNOVATION SPACE

Learning

Figure 1a shows the hypothesized learning space. When people are high on the efficiency dimension, they can rapidly retrieve and accurately apply appropriate knowledge and skills to complete a routine task or solve a familiar problem. Typically, learning scientists use measures of speed, accuracy, and consistency to capture efficiency. Everyday examples of efficiency include people who have a lot of experience with

certain types of tasks and problems; for example, doctors who have frequently performed a specific surgery or skilled typists. Efficiency is important in all domains. As a field, we have learned much about how to accelerate development along the efficiency axis through properly organized and situated practice (Bransford, Brown, & Cocking, 1999).

While efficiency works well when people operate in constant environments, there are perils to efficiency. Experimental studies show that efficiency can produce "functionally fixed" behaviors where people perseverate on previously efficient schemas instead of letting go to see new alternatives (Luchins, 1942). For example, nine-year-old children trying to learn about fractions often count pieces by using efficient, whole number counting schemas, which interferes with their abilities to interpret the pieces as parts of wholes (Martin & Schwartz, in press). Hatano and Inagaki (1986) discuss "routine experts" who become increasingly efficient at solving familiar problems, but who do not engage new problems and situations.

Innovation involves creating new skills and concepts, often as a way to adapt to new situations. As a field, we know less about innovation. Relevant psychological variables, like creativity, are often considered general traits or skills. Efforts to cultivate creativity typically involve brainstorming techniques. "Content-lite" approaches like these are not ideal, because meaningful innovation requires integration with an extensive body of efficient knowledge. Ericsson, Krampe, and Tesch-Römer (1993), for example, found that original intellectual contributions to a field occur after people spend ten years developing the requisite domain expertise.

Ideally, the goal of education is to position people in the upper-right corner of Figure 1a. People often think of innovation and efficiency as incompatible (e.g.,

discovery learning versus "back to basics"). However, the possibility of balancing efficiency and innovation is highlighted by Hatano and Inagaki's (1986) notion of adaptive experts who have a wealth of efficient knowledge but are also able to adapt to situations that require innovation. Efficiency in some processes (decoding written words) frees attention for other things (reading for meaning). If people have solved aspects of a complex problem before, this helps make sub-problems routine, freeing them to concentrate on other aspects of the situation that may require innovation. At the same time, it is important to resist practiced responses to situations that do not call for them. A major challenge for the learning sciences is to understand how to balance efficiency and innovation in learning.

Design research can contribute by developing interventions that strike the right balance of innovation and efficiency experiences and that place students on a trajectory towards adaptive expertise. This research should also create ways to determine whether students are on that trajectory, and this requires more than borrowing standard off-the-shelf measures of efficiency. Efficiency measures can misdiagnose the value of an innovative experience. For example, Schwartz and Bransford (1998) asked college students to innovate representations of data from classic studies of memory. Other students wrote a summary of a parallel chapter. On a subsequent true-false test – a standard test of efficient factual recall – the summarize students did better than the innovate students. However, a new type of measure revealed the value of the innovative experiences. Students from both conditions heard a common lecture that reviewed the studies and their implications for human behavior. A week later, the students had to predict the results of a new, but relevant, experiment. The innovation students produced

twice as many correct predictions with no increase in wrong predictions – they had been prepared to learn from the lecture and adapt what they learned to predict the results of the new experiment. The benefit of the innovative experiences was missed by the standard efficiency-oriented assessment, but it was captured by a measure that looked at students' subsequent abilities to learn. As a field, it is important to develop ways to measure the benefits of innovative experiences, lest these experiences look useless on standard tests of efficiency.

Science

We can export the learning space to scientific inquiry without too much damage. Efficiency research nails down the regularities identified by prior work. It involves applying, refining, or testing prior beliefs. One example would be the human genome project; the techniques of gene sequencing were already established and finishing was a matter of time. Another example occurs when well-delineated hypotheses create testable predictions, and at their finest, can create a "Galilean experiment" that adjudicates between two theories (Medawar, 1979).

As with learning, there are scientific perils to an efficiency-only approach. Recent political developments in educational research push for large clinical studies to determine which models of instruction are the most effective. As the Design-Based Research Collective (2003) noted, "the use of randomized trials may hinder innovation studies by prematurely judging the efficacy of an intervention" (p. 6). The political push for clinical trials may promote the use of efficient experimental methods to make choices between sub-optimal alternatives. Moreover, good research designs may be compromised by inappropriate measures, such as evaluating students' efficiency in

school tasks rather than their potential for future learning and decision making beyond school.

Innovations in science take many forms. One form involves explanation; for example, why children find division harder than multiplication. Another form involves the discovery of a new phenomenon like X-rays. In contrast to efficiency research, innovative science does not necessarily depend on generalization or causal identification. A single case is often sufficient to challenge a long standing theory. The invention of the telescope led to the discovery of lunar imperfections, which undermined the prevailing theory of heavenly perfection (Kuhn, 1957).

There are also perils to innovation-only research. In the context of design experiments, Brown (1992) writes, "It is not sufficient to argue that a reasonable endpoint is an existence proof, although this is indeed an important first step" (p. 171). One peril is that because the work is about innovation, it often needs to let go of current theories. This can create a tower of innovation babble with little short-term hope of cumulative knowledge. A second peril is that if innovations must stand on their own, with limited support from prior theory, the research is difficult and runs a high risk of failure. diSessa and Cobb (2004), for example, argue that a preeminent goal of design experiments is to create new theories. This may be a fine goal for brilliant researchers at the top of their game, but for the rest of us, it is a recipe for heart-felt platitudes.

Innovation-only and efficiency-only approaches are not sufficient for the types of progress needed to improve education. The ultimate goal of research is to contribute to a growing body of knowledge that comprises tested "truths" but adapts to new findings and historical times. The challenge for design experiments is to find a way to balance the

goal of innovation with the need for subsequent efficiency. We propose that a focus on instrumentation can help achieve this balance.

INSTRUMENTATION FOR PROMOTING INNOVATION AND EFFICIENCY

If it were possible to quantify contributions to the advancement of science, instrumentation would compete well. New instruments open territories that scientists quickly populate. One only needs to look at the effects of fMRI on psychology. Interestingly, instrumental innovations are often independent of research methodology. Videotaping, for example, can be used in clinical, experimental, and cultural applications. Sometimes we wonder if debates over research methods are missing the action. The most highly cited authors, at least in psychology, are those who make new instruments for research. Here, we describe examples of how instrumentation research supports innovation and the subsequent turn to efficiency. We begin with science, and then develop the parallel for individual learning.

Science

Innovation in Instrumentation

New instruments often foster scientific innovation by enabling scientists to see what they could not see before; cell stains, telescopes, and the habituation paradigm are just three examples. They exemplify the first half of the instrument equation – the "apparatus" that makes phenomena observable. Passive apparatus (cameras) and emissive apparatus (radar) are staples of the natural sciences. In the behavioral sciences, researchers often use perturbing apparatus that trigger processes to make their features more visible. For example, psychologists can use an "apparatus" of word stimuli to trigger people's thoughts and see how they affect memory.

Design experiments, because they are interventions, can also be recast as a

"perturbing" apparatus. Cobb et al., (2003) state, "Prototypically, design experiments entail both 'engineering' particular forms of learning and systematically studying those forms of learning…" (p. 9). When design researchers devise novel lessons, technologies, or social practices, they are designing a new apparatus for perturbing the environment to reveal processes of learning. Ideally, the apparatus can also be reused, if the resulting learning processes are desirable.

The second half of the instrument equation is the development of measurement. Measurement converts observations into precise communicable information. Though measurement reduces the totality of a phenomenon, the gains in precision can aid innovation. Precise measures can pick up aberrations from the expected. Astronomers in the early 1800's found that the measured positions of Uranus did not match its predicted orbit. This anomaly led to the hypothesis and eventual discovery of an eighth planet, Neptune. Galileo had seen Neptune through his telescope, but he observed it as a star. The precision of measurement, and not pure observation, led to discovery.

Taking the Turn toward Efficiency

Instruments that were once innovative may subsequently support efficient progress in science. Piaget created instruments to evaluate children's cognitive function. The instruments themselves could be disseminated and evaluated independently of Piaget (e.g., cross-cultural applications). This allowed the research community to take a turn from an innovative but singular set of studies to a more efficient mode of research that refined the instruments and addressed what causes change in the measurements (and whether Piaget's theory was correct).

The measurement component of instrumentation permits others to determine if

they are "seeing" the same thing. Measurements can be quantitative, for example, the time to complete a task. Measurements can also be qualitative, for example, a set of observational criteria for the presence of a phenomenon. If researchers want, they can convert qualitative measurements into a quantitative form by recording the frequency or intensity of an observation. The advantage of quantification is that it permits researchers to use the substantial structural apparatus provided by mathematics to draw inferences. However, quantification is not a prerequisite of measurement, and oftentimes it is a mistake to force a set of observations into a particular mathematical model (e.g., a linear model).

A challenge for innovation research is that "the decision to employ a particular piece of apparatus and to use it in a particular way carries an assumption that only certain sorts of circumstances will arise" (p. 59; Kuhn, 1970). Some researchers reject the idea of using measures because they worry the measures will foreclose the possibility of detecting the unanticipated. Consequently, many rely on narratives rather than discrete measures to create inter-subjectivity. An extreme position, like that of Eisner (2001), argues that a research report should help the reader <u>experience</u> what the researchers saw, including their epiphanies and failures. Unlike some traditionalists (Shavelson et al., 2003), we do not have a strong opinion about narrative. We do not know of any evidence one way or another that determines whether narrative yields precise agreement between the researcher and the audience.

Personally, we report discrete process and outcome measures in scientific papers. This does not mean that we are not highly attentive to occurrences that elude our measures. (We are present throughout our studies, we videotape, and we collect

artifacts.) It is naïve to assume that researchers who use measures are not also keen on discovering processes and outcomes they never imagined. For example, when we design a new paper-and-pencil assessment, we exhaustively code every response looking for interesting patterns. In the studies below, each instrument averaged seven different categories of response. We would love to report all the responses and patterns that we find. However, like all scientific authors, we decide which of the patterns will be most compelling and economical to report – this is often a simple percent correct, but not always.

Implications for Design Experiments

A place for substantial improvement in design research involves the use of measurement. Design research is quite good at developing techniques for the apparatus half of the equation – innovative instruments that precipitate effects. However, most design research has not finished the equation by developing innovative measures suited to those effects. This lack of measure is surprising. Unlike ethnographers, design researchers are orchestrating "what could be" rather than observing what already exists, and therefore, they have must have some goal in mind. Ideally, this goal would be specific enough that it is possible to begin precisely measuring its attainment.

One hope of design research seems to be that the instructional apparatus will jump from the quadrant of high-innovation and low-efficiency in Figure 1b to the quadrant of cumulative knowledge, perhaps through large-scale clinical trials involving random assignment and standard achievement measures. Most standardized tests of achievement and intelligence, however, are created to rank people and not precisely reveal the component knowledge and processes responsible for an answer. Without working on

measures that are tightly matched to the perturbing intervention, the research will yield claims like, "Our intervention was *significantly better* than standard practice." Though good at satisfying evidence-based educational policy, we fear the vagueness of the measure will not contribute to the cumulative growth of scientific knowledge. For example, when the social climate changes what it means to "do better," these studies will become irrelevant instead of leaving documentation of exactly what learning a particular design yields or how to improve it. Creating measures that are tightly coupled to an apparatus of change can better facilitate a productive turn to efficiency.

Learning

To us, it is clear that instrumentation can support innovation and the turn to efficiency in science. The idea that working on instrumentation can also propel individual learning is less obvious. Measurement, in particular, often conjures images of students mechanically using rulers. This is an efficiency-only take on measurement that presupposes the techniques and outcome categories already exist in the mind of the learner. The innovation side of measurement is not so banal, though it has been overlooked in the research literature. The standard cognitive study of scientific reasoning emphasizes methodological thinking over measures. People receive a set of well-defined input and output variables (the instruments), and their task is to design unconfounded experiments to discover the relations (Chen & Klahr, 1999; Kuhn, Schauble, & Garcia-Mila, 1992). Our experience with hundreds of adults is that it is not the ability to reason within an experimental design that is the difficult part of scientific thinking. Novices quickly learn about confounds, though they sometimes forget to use this knowledge efficiently or find it tedious. The more difficult challenge is developing measures suited

to a specific domain. By the same token, asking novices to attempt to innovate measurements can be an excellent source of domain learning. Creating measurements encourages specificity in understanding. For example, asking students to measure how well people learn from a reading passage can help them develop more differentiated knowledge of what it means to learn, such as whether it is more appropriate to measure free recall, recognition, comprehension, inference, or transfer.

Another benefit is that measurement can indicate places that require innovation. As a rather grand example, we consider Plato's learning paradox. This paradox raises doubts about whether people can innovate new knowledge, and accepting the paradox leads to an efficiency view of learning that emphasizes the refinement of prior knowledge (e.g., innate concepts, language modules, phenomenological primitives, and so forth). Through the dialog of the Meno, Plato (1961) formulates two components of the paradox:

> But how will you look for something when you don't in the least know what it is? How on earth are you going to set up something you don't know as the object of your search? To put it another way, even if you come right up against it, how will you know that what you have found is the thing you didn't know? (80.d).

The first half of the paradox asks how people can look for knowledge if they do not already know what they are looking for. Plato's solution is that incommensurables alert people to the need for innovation. The term incommensurable refers to the situation where multiple elements cannot be measured within the same rational system. For example, if we try to determine whether an Olympic weight lifter broke the world record by more than a long jumper did, we cannot use weight to measure distance or distance to measure weight – the performances are incommensurable. Thus, a failure in the

measurement system lets one know where to look for new knowledge. It causes the disequilibrium that begins the search for resolution.

The second half of the paradox asks how can people recognize whether they have found knowledge if they do not already know it. The solution is that people know they have found new knowledge when the incommensurables can be explained within the same system. In the case of weight lifting and long jump performances, one makes them commensurable by using standardized scores. People know they have learned something new, because it is possible to relate what they could not previously. We return to the example of standardized scores below, but for our present purposes, it is noteworthy that Plato resolves the learning paradox by offering measurement as a premiere example of an innovation in learning.

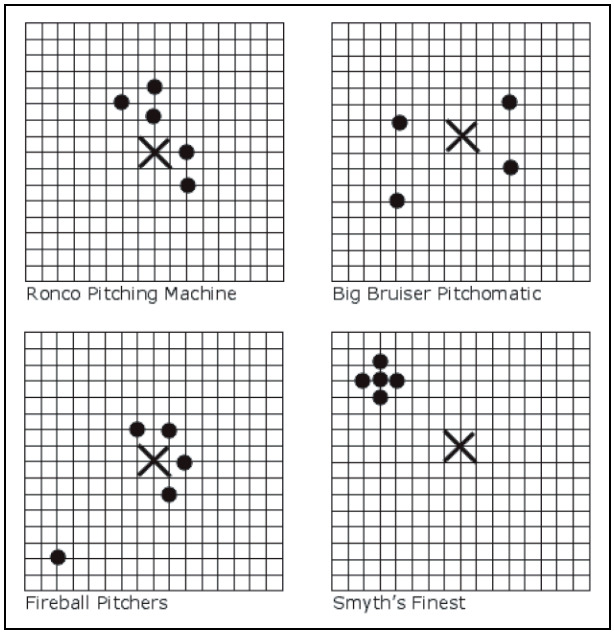

Figure 2. Students innovate a reliability index to measure pitching machines. (Schwartz & Martin, 2004).

In our work teaching children statistics, we capitalize on the potential of measurement for propelling learning. An example from the materials we use to teach variability can help demonstrate the point. Each grid in Figure 2 shows the result of a test using a different baseball-pitching machine. The black circles represent where a pitch landed; the X is the target. Students receive the innovation task of developing a formula or procedure that computes a reliability index for each machine to help shoppers.

The pitching grids were designed to help students develop more differentiated and structured knowledge. By highlighting key quantitative distinctions, the contrasting grids alert learners to the properties their measurements need to reconcile. For example, most students initially misinterpret variability as a lack of accuracy. The pitching grids specifically include an example where all the pitches are extremely close together, yet they are far from the target. This helps the students notice that variability and lack of accuracy should be differentiated.

By asking students to innovate a single measure by which to compare the machines (the reliability index), this task also promotes a more structured understanding of variability, because the students' formula must accommodate the various dimensions along which the grids differ. For example, the grids use different sample sizes. Many students begin by summing the distances of the pitches from the target, but they quickly realize that grids with more pitches will tend to get higher variability scores, even if the pitches are close to the target. A simple summation measure makes samples of different size incommensurable. The need to handle sample size becomes a structural element of students' understanding of variability.

We do not expect students to innovate the conventional measurement. Instead,

our assumption is that the innovation activities prepare the students to understand efficient expert solutions more deeply. For example, when they hear the standard solution for finding variability, they will appreciate how dividing by 'n' elegantly solves the problem of comparing samples of different sizes (by taking the average of the deviations from the mean). As fits the learning space, we have students accelerate on the innovation dimension first, before they take the turn to efficiency. An emphasis on instrumentation, in this case measurement, can facilitate this trajectory. An alternative would be to just tell the students how to compute variance at the outset. We believe this yields efficiency, but it does not create a trajectory towards adaptive expertise. The following section tests this belief.

A DOUBLE DEMONSTRATION OF THE INNOVATION-EFFICIENCY SPACE

Figure 3 summarizes our claims so far. We believe that both scientific progress and individual learning benefit from accelerating first on the innovation dimension before taking a turn to efficiency, and we propose that this trajectory is particularly well-supported by an effort to develop instrumentation. To make our claims more concrete we provide an example of this double-trajectory in a study we did with 9th-graders. Notably, none of the research we describe is about proving causes. Instead, it is about demonstrating the discriminant and ecological validity of our instructional apparatus and measures, which we believe is one place where design research can excel in contributing to scientific knowledge.

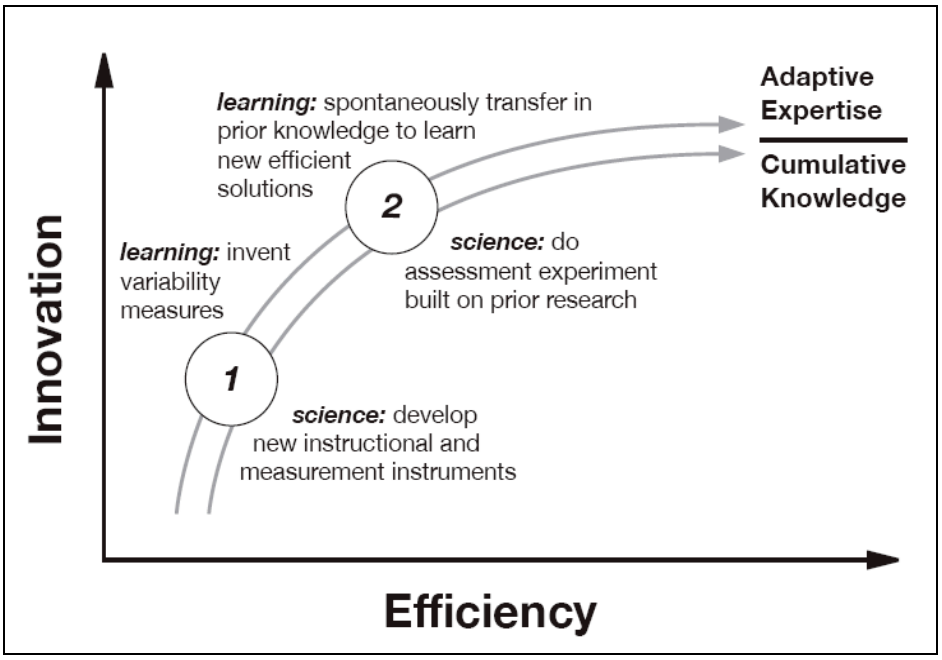

Figure 3. Taking the turn in learning and science.

Demonstration 1: High on Innovation, Low on Efficiency

This research involved six classes of 9th-grade algebra students. It was the students' first introduction to the notion of innovating measures and the topic of variability. It was our first effort at building a new apparatus to help students innovate on the topic of variability and at building new measures to detect the effects. The pitching-grids of Figure 2 provide an example of one instructional apparatus. We provide an example of an innovative measure below. A complete description of the full set of instruments and results may be found in Schwartz and Martin (2004).

Students spent a few hours innovating their own ways to measure variability in a variety of tasks. Students worked in groups, and there were class discussions about the invented solutions. Students never invented a canonical solution, and the teacher did not present one during the innovation phase. However, at the end of the innovation activities,

the teacher gave a 5-min lecture on the mean deviation (an efficient way to measure variability) and students practiced for another 10 minutes. The question was whether students' innovation activities would prepare them to learn the efficient solution from the brief lecture and practice, and whether we could measure any special benefits of the innovation activity.

All six classes received the same instruction. Putting classes into control conditions was premature. Still there were ways to make scientific progress. We included techniques, such as benchmarking our instruments, to support plausible inferences on whether we (and the students) were on a good trajectory. By plausible inference, we mean the evidence confirms our expectations. This is a much weaker warrant than the standard of falsification in efficient science, and these kinds of designs cannot guarantee that there were not other sources of influence. However, in early stages of research they can be a valuable source of preliminary evidence. We have conducted many design experiments where the results did not support plausible inference, leaving us to decide whether there was a problem with our instruments or whether our guiding ideas were wrong. Deciding whether to try again is a problem faced by all researchers. Instrumentation cannot guarantee successful innovation – nothing can. Therefore, we usually "bet low" by conducting small studies, and then pursue the most promising results.

The most important of our techniques for plausible inference is "targeted measurement." We try to tune our measures to specific features of our intervention. For example, as part of our assessment, we used a number of off-the-shelf techniques for measuring efficiency, such as asking students to compute variability and to solve word

problems. Students improved from 5% at pretest to 86% at posttest. A year later without any intervening review, they were still at 57%. This measure indicated the students had learned the efficient solution, which is very important, but we thought it missed the unique benefits of the innovation activities. We needed to invent new measures that targeted the novel benefits of our instruction.

Table 1. Percentage of students who successfully explained why a formula uses a given operation.

	9th-Graders		College Students	
	Time of Test		Semesters of Statistics	
	Pretest	Posttest	None	One
Why does $\Sigma \lvert \bar{X}\text{-}x \rvert / n$ divide by n?	6%	63%	0%	12%
Why does $m = y_2\text{-}y_1/x_2\text{-}x_1$ subtract x_1?	10%	14%	11%	29%

We included a large number of targeted measures specifically addressing our expectation that innovating measures would help students subsequently understand why an efficient measurement procedure does what it does. One example comes from our expectation that students who worked with the pitching grids would realize that their measurements needed to handle the different sample sizes in different grids. We thought this would prepare them to appreciate how canonical variability formulas accomplish this task. To test this idea, we developed a "symbolic insight" measurement. Students receive a formula and have to explain one of its symbolic operations; for example, "Why does this variability formula divide by 'n'?" As a comparison, we also created symbolic insight questions about another formula they had learned recently but not through our innovation-to-efficiency curriculum. We exhaustively coded the different types of

answers and found a number of interesting patterns (see Schwartz and Martin, 2004). Table 1 simplifies the data by using an aggregated measure of symbolic insight. The students did better with the variability formula than other formulas they had been taught in other ways, and they showed greater gains than a benchmark of college students who had taken a full semester of college statistics. This leaves us with the plausible inference that our intervention helped students develop symbolic insight, and that this insight is not a common outcome of other forms of instruction.

Without a control group, we cannot go beyond the plausible inference that the innovation component of the lessons prepared students to learn the variability formula so well from the brief lecture and develop symbolic insight. Being early on the innovation curve, the time was not right for efficient tests of causal hypotheses. However, the study equipped us with the instrumentation to find out.

Demonstration 2: Taking the Turn

The second demonstration involves the same students after two weeks of innovating statistical measures and subsequently hearing efficient solutions. The students were further along the curve in their statistical knowledge, and they were able to show a turn towards adaptive expertise. The demonstration also shows how our research was able to take the turn to efficiency through the use of an assessment experiment. To better explicate the design and purpose of the experiment, we provide some background.

The experiment arose from a concern that most current assessments of knowledge use sequestered problem solving (Bransford & Schwartz, 1999). Like members of a jury, students are shielded from contaminating sources of information that might help them learn during the test. It appears that this assessment paradigm has created a self-

reinforcing loop where educators use efficiency-driven methods of procedural and mnemonic instruction that improve student efficiency on sequestered tests. However, we suppose that a goal of secondary instruction is to put students on a trajectory towards adaptive expertise so they can continue to learn and make decisions on their own. Like many others, we fear that measures of the wrong outcomes drive instruction the wrong way. In prior work, we had shown that preparation for future learning (PFL) assessments, which directly examine students' abilities to learn, are a viable alternative to sequestered assessments, and they better reveal the strengths and limitations of different forms of instruction in college-level psychology (Schwartz & Bransford, 1998). However, one series of studies with one demographic profile and in one content area is insufficient. Moreover, in those studies, students were directly told what they were supposed to learn. Ideally, good instruction can help students learn in the future without explicit directives. Thus, with the current assessment experiment, we wanted to continue work on PFL measurements using a new age group, a new topic, and a new format that determined whether students spontaneously took advantage of a potential learning resource.

The assessment experiment crossed the apparatus of instruction with the method of measurement. We first describe the two instructional conditions for learning about normalizing data shown at the top of Figure 4. Students received raw data that required them to compare individuals from different distributions to see who did better. For example, the students had to decide if Bill broke the high-jump world record more than Joe broke the weight-lifting record given data of the top jumps and lifts that year. Three randomly selected classes were assigned to the invention condition. These students had

to innovate their own way to solve this problem. There were neither class presentations nor feedback, isolating the value of innovation from other features in the larger design experiment. The other three classes received the tell-and-copy treatment. These students were taught an efficient visual procedure, which they copied using the data sets.

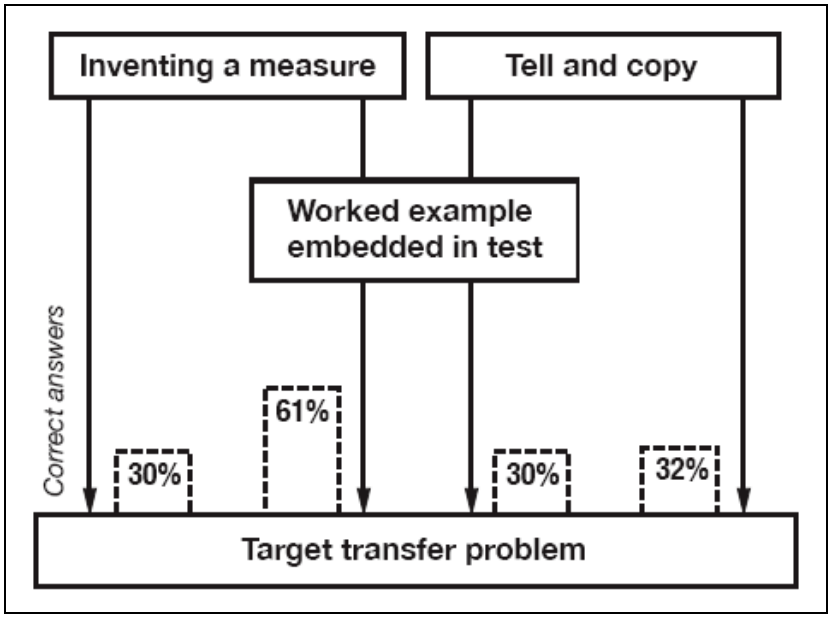

Figure 4. Design of assessment experiment and results. (Schwartz & Martin, 2004).

The second factor involved the method of measurement and whether students received an embedded learning resource in their posttest several days later. In Figure 4, this is shown by whether arrows go through the middle box. The resource was a worked example that showed how to compute standardized scores (see Appendix). The example showed how Cheryl determined if she was better at the high dive or low dive. The students had to follow the example to determine if Jack was better at high jump or javelin. Half of the students from each condition received this worked example as part of their posttest and nearly everyone followed it correctly. The question was whether students followed their usual efficiency-oriented practice of treating the example as a

"plug and chug" problem, or whether they adapted to learn what the problem had to offer. To detect spontaneous learning, there was a target transfer problem later in everybody's posttest (see Appendix). It involved a different context and format, and its solution depended on using standardized scores as in the worked example. The worked example and the target transfer problem created a novel *double transfer assessment* of preparation for future learning; students needed to transfer in to learn from the worked example and they needed to transfer out from the worked example to solve the target transfer problem.

Figure 4 shows the percent of students who solved the target transfer problem. Students who had innovated their own methods for normalizing data learned the efficient solution from the embedded worked example and spontaneously transferred this learning to solve a novel problem, more so than students who had been told and had practiced a specific visual technique for normalizing data. This difference shows the effectiveness of innovation activities in helping students take the turn towards adaptive expertise. Students in the innovation condition who did not receive an embedded resource were probably still better prepared to take the turn than the tell-and-copy students, but they had no opportunity to demonstrate that readiness and looked the same. Thus, the PFL measure demonstrated discriminant validity, because it detected a difference that was missed when there was no embedded learning resource. This example also shows how PFL assessments can be sensitive measures of levels of understanding that we care about but that can be missed by sequestered measures of efficient problem solving.

There are two points to make with this example. One point is that student learning can advance on a curve that first accelerates along innovation and then turns to efficiency. The value of innovation activities for student learning seems obvious in

retrospect, but there is a great deal of confusion in this area (see Schwartz & Martin, 2004). We hope these studies are a useful demonstration of the hidden value of innovation for subsequent efficiency.

The second point is that instrumentation can help innovative research take the turn to efficiency. For example, in the study above, we taught the classes ourselves. We then gave our instruments to classroom teachers who implemented the study the following year with seven new classes. Results indicated a very high degree of replication.

Good instrumentation research does not have to depend on causal claims, and good instrumentation does not have to inhibit the innovative spirit of design experiments. Nowhere in this work did we isolate specific causal ingredients, and we were able to leverage our instrumentation with little cost to the design experiment itself. This instrumentation focus permitted us to make cumulative and replicable progress without cutting the research into expensive experimental designs that are efficient in terms of causal proof but inefficient in terms of resources.

WARRANTS OF INNOVATION

If anything unifies the literature on design experiments, it appears to be the goal of useful innovation. Until now we have skirted a fundamental question – what is an innovation? How can a researcher or funding agency decide if there is productive movement on the innovation dimension? It would be useful to have some grounds for identifying and evaluating design innovations, especially if that is what design researchers claim to do! The goal here is not a set of practical guides to innovation (e.g., instrumentation) or practical agreements of innovation. There are already a number of practical criteria for judging innovation. For example, the legal system judges whether an innovation deserves a patent based on categories including novelty, non-obviousness,

and utility. Historically, however, there was so much variability in legal opinion that it was necessary to make a single federal court to handle patent cases. Ideally, we can find a logic for evaluating innovation that is more stable than pragmatic agreement.

Our goal here is to provide a "sacrificial first draft" for a discussion of warrants for scientific innovation. We think this is important because design researchers can best justify their work in terms of innovation. If so, it is important to be clear about one's grounds for making an innovation warrant. The significance of being clear comes from the story of a university job talk. The candidate described a few heroic teachers who had teamed together to innovate a successful program in an environment of indifference and poverty. This candidate was making an innovation argument by showing that what people implicitly thought was impossible, actually could be. The faculty in the audience, however, challenged the work. The candidate made a mistake and started to defend causal claims about the teachers' success. This single case could never crisply defend causal claims. The work came off as speculation and craft knowledge. We wish the candidate had said, "Those are great questions, and the value of this single data point is to show that these are important questions precisely because our theories should, but cannot, explain it. Here are some of the instruments I developed to help us move forward…." As in all things, knowing the logic of one's claims can only help.

As we began thinking about the problem of evaluating innovations, we realized that we, and perhaps others, had confounded two distinct forms of innovation – innovating new knowledge and innovating material change. Producing new knowledge about learning is different from producing a new artifact or social practice, though they are both important innovations. The difference between innovating knowledge and

material change can be captured by two possible stances on the value of design research that we label "scientific" and "substantive." Innovations in scientific design involve discovery and theory, whereas substantive innovations involve changes to the circumstances of learning. We begin by drawing the distinction between the two, and then we describe one warrant for innovation that has logical force for both types of design research.

Scientific and Substantive Design

Scientific design proposes that design methodologies are the best way to develop innovative knowledge in educational research. The assumption is that there are truths about the world that need to be uncovered. diSessa and Cobb (2004), for example, propose that a significant goal of design research is "ontological innovation"—the invention of new scientific "categories that do useful work in generating, selecting among, and assessing design alternatives." (p. 78). The goal is to uncover categories of phenomena that support explanation. We suppose these authors actually meant "epistemic innovation," because ontology refers to what exists and epistemology refers to knowledge of what exists. These authors are not proposing that the value of design experiments is to create new existences, but rather to create new knowledge.

In contrast, substantive design holds that the goal of a design experiment is to improve education per se (as opposed to just improve knowledge of education). The goal is to make changes to the world, and design is the shortest distance from idea to change. Cobb, Confrey, diSessa, Lehrer, and Schauble (2003) state that one purpose of design experiments "…is to investigate the possibilities for educational improvement by bringing about new forms of learning in order to study them" (p. 10). These authors

appear to advocate the creation of new forms of learning, but they justify the endeavor from a scientific design stance. A pure substantive position does not justify its efforts in terms of developing scientific knowledge. Some of the best technology innovators appear to adopt a strong substantive position. They create stunningly innovative technologies to support learning. They are less committed to accounts of why or whether these technologies promote learning. This seems a useful way to make progress, though it is important to appreciate that the goal of educational design research is not technological innovation per se, but rather innovation in learning practices. Thus, pointing to an innovative technology is less compelling than pointing to an innovative learning practice it creates.

Substantive design is appropriate to intervention research because it holds that research can change reality rather than just study it. G. H. Mead (1899) captures the quixotic implication:

> In society, we are the forces that are being investigated, and if we advance beyond the mere description of the phenomena of the social world to the attempt at reform, we seem to involve the possibility of changing what at the same time we assume to be necessarily fixed. (p. 370)

The idea that design can innovate new forces and facts by which the world operates is inconceivable in domains like physics. But human affairs take place in a social world, and the laws that regulate social behavior have their own organization which is consistent with, but under-determined by, physical and biological laws. Therefore it may be possible to change the social world by design (e.g., Searle, 1995; Simon, 1996). Freire (1970) refers to this as the dialectic of objective and subjective

experience. For example, a capitalist economy has a different set of forces than a communist one. At some reductionist level, people in capitalist and communist societies are built from the same fundamental material laws, but at the level of subjective experience and the objective analysis of that experience, they operate according to different rules. Karl Marx's theory built on the possibility of changing the forces that regulate our lives. When he wrote, there were no full-blown communist states. His theory could only be true to the extent that it could change reality to fit itself. This is the notion of praxis, where the proof of a theory is in the change it creates (Cook, 1994). Praxis is highly relevant to claims that a new plan for classroom organization will change what students learn, and it provides an interesting way to warrant a substantive design innovation.

A Warrant for Innovation

Developing warrants for useful innovation is important, lest the mechanisms for evaluating innovative work reduce to consumerism. Toulmin (1972) provides a useful list of types of conceptual change in science: (1) extension of current procedures to fresh phenomena; (2) improvement in techniques for dealing with familiar phenomena; (3) intra-disciplinary integration of techniques; (4) inter-disciplinary integration of techniques; and, (5) resolution of conflicts between scientific and extra-scientific ideas. These can be recast into warrants for innovation. For instance, (1) an instrument can be considered an innovation if it permits us to observe a fresh phenomenon. Unlike efficient science that gains warrants through reliable prediction of the future, each of the resulting innovation warrants would depend on showing a difference from the past.

Toulmin's first two categories suggest that one method of warranting an

innovation is to show that was has been innovated was previously absent. This works well for scientific design, because there is a documented canon of knowledge one can exhaustively search to demonstrate prior absence. However, demonstrating absence is problematic for substantive design in the social sciences. A design may create a seemingly novel social practice, but then it may turn out there is a tribe in a remote location that already engages those practices.

Toulmin's latter three conditions, which emphasize integration, suggest a warrant that can work for both scientific and substantive design. The warrant is the reconciliation of incommensurables, and returns us to Plato's resolution of Meno's learning paradox. In this case, an innovation warrant depends on finding an incommensurability or contradiction and then reconciling the horns of the dilemma in a synthesis. The synthesis is an innovation by definition, because it resolves what extant knowledge systems could not. For learning, one example is the students in our study trying to relate high jumping and weight lifting scores. They could not measure distance in terms of weight, or weight in terms of distance. The innovation they were struggling towards was a way to normalize data so they could compare unlike measures. For them, the solution of standardized scores was an innovation on logical grounds, because it reconciles within one structure what their prior knowledge of absolute magnitudes could never do. Similarly, in science, a powerful logical warrant for innovations in knowledge occurs when the innovation provides a way to put previously incompatible evidence or theory in the same rational structure. Galileo was a master at generating evidence that contradicted the theories of the day, and then presenting his alternative that was able to synthesize all the evidence within one framework.

The reconciliation warrant for innovation can also work for substantive design. In this case, it involves the reconciliation of contradictory forces rather than contradictory knowledge. The physical world does not contain any contradictions (nature might contradict our theories, but it cannot contradict itself), but, in the social world of goals and means, contradictions are possible. In remote Alaskan villages, for example, schools attempt to improve village life by preparing native students for jobs that require leaving the village for the city. A substantive design that removed this contradiction would have a strong warrant for innovation through praxis.

To achieve this warrant in substantive design, it is necessary to first identify an incommensurability or contradiction. This is a needs assessment, but one with logical force. Saying that children have confused identities is a weak needs assessment compared to one that identifies the contradictory forces that lead to confused identities. Engestrom (1994) takes this needs assessment approach, for example, by identifying contradictions in social organizations (e.g., the elderly health care system), and then creating innovations that resolve those contradictions (e.g., postal workers check elderly on their rounds). We find his methodology compelling because it illuminates the contradiction and shows how the innovation attempts to reconcile this contradiction, whether or not it works.

In our research, we identified a contradiction that we tried to reconcile. Most educators want their students to be on a trajectory to adaptive expertise so they can continue to learn and make their own choices as citizens. At the same time, educators try to promote this goal by implementing efficiency-only curricula and assessments, which we believe unwittingly contradicts the goal of adaptive expertise. We tried to resolve this

contradiction by innovating a knowledge framework that puts innovation and efficiency together rather than as opposites, and we innovated a pedagogy in which innovation and efficiency practices can co-exist and put learners on a trajectory towards adaptive expertise. Ultimately, we believe our substantive design efforts fell short. The failure was not in the study's outcome; we showed that efficiency measures can miss the value of innovative experiences whereas PFL measures do not. Rather, the failure was in proving the contradiction: perhaps efficiency-only training does lead to adaptive expertise in the long run and there is no contradiction. We hope that our instrumentation focus will enable us to generate subsequent research to determine whether, or when, we are correct.

CONCLUSION

We want to make design experiments a more productive scientific endeavor. Innovating in educational settings is largely intractable by standards of efficient science; it is too costly to use the sample sizes and control conditions needed to test every idea of what could be. At the same time, "trying stuff out" is not adequate either. We have been frustrated by the design experiment debate, because it has reified methodological positions as incommensurable, while ignoring those things that are most important to working scientists. Though discussions of method and theory are very important, empirical scientists in most fields spend their time working on instrumentation. We proposed that it might be profitable to position design experiments in a larger space of innovation and efficiency. The question is, what features might be added to design research to ensure it maximizes the chances for innovation while also setting the stage for more standard tests of scientific value.

We argued that a focus on instruments that both precipitate and measure effects has historically been effective at supporting innovation and the turn to efficiency. There

is a repertoire of ideas for evaluating instrumentation that do not depend on identifying causality or generality (e.g., discriminant validity). We also provided an empirical demonstration where innovating measures led to impressive learning gains for students, and hopefully demonstrated the potential of our PFL measures for advancing science.

In the last part of the chapter, we initiated a discussion around the logic of justification rather than the practical process of innovation. There can be no logical method for guaranteeing discovery or innovation (Popper, 1968; Phillips & Burbles, 2000), but we thought it might be possible to work towards a logic for warrants of innovation. We think it is important for the design experiment community to create compelling standards for evaluating its success at innovation. We have found that people often try to use efficiency arguments that cannot succeed, or they provide no warrants at all. We tried to clarify two arguments for the value of design innovations – scientific innovations that involve knowledge and substantive innovations that involve change. We think the substantive design position is particularly relevant to design researchers, but its logic of justification has not been sufficiently explored.

We argued that a logical warrant for innovation is the resolution of incommensurables. (Not surprisingly, this is also the type of innovation we asked students to pursue in reconciling contrasting cases.) This warrant helps draw a strong distinction between the logic of efficiency and innovation. Whereas criteria of efficiency depend on predicting future regularities, the criteria of innovation depend on reconciling past irregularities. Reconciling past irregularities requires planned design, but the causal components of the plan are not being put to test. What is being put to test is whether the irregularity is necessary or whether it can be resolved.

We find the two goals of design research – discover knowledge versus plan change – equally compelling. It is an empirical question whether designing our way to a better educational system is more effective than first developing scientific knowledge and then engineering change from established "laws." We do not believe these approaches are incompatible, and in fact, we suspect that both are needed. This consideration led, in part, to our proposal that design experiments would be well-served by explicitly engaging in instrumental innovation that paves the way for efficient scientific methods, while also providing the apparatus for creating and recreating the qualities of what could be.

REFERENCES

Barron, B. J., Schwartz, D. L., Vye, N. J., Moore, A., Petrosino, A., Zech, L., Bransford, J. D., & CTGV. (1998). Doing with understanding: Lessons from research on problem- and project-based learning. Journal of the Learning Sciences, 7, 271-312.

Bransford, J. D. & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad and P. D. Pearson (Eds.), Review of Research in Education, 24, 61-100. Washington, D.C.: American Educational Research Association.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). How people learn: Brain, mind, experience, and school. Washington, DC: National Academy Press.

Brown, A. L. (1992). Design experiments: theoretical and methodological challenges in creating complex interventions in classroom settings. Journal of the Learning Sciences, 2, 141-178.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. Child Development, 70, 1098-1120.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. Educational Researcher, 32(1), 9-13.

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design Research: Theoretical and Methodological Issues. The Journal of the Learning Sciences, 13, 15-42.

Cook, T. E. (1994). Criteria of social scientific knowledge: Interpretation, prediction, praxis. Lanham, Maryland: Rowman & Littlefield.

Design-Based Research Collective (2003). Design-based research: An emerging

paradigm for educational inquiry. Educational Researcher, 32(1), 5-8.

diSessa, A. A., Cobb, P. (2004). Ontological innovation and the role of theory in design experiments. Journal of the Learning Sciences, 13, 77-103.

Eisner, E. W. (2001). Concerns and aspirations for qualitative research in the new millennium. Qualitative Research. 1, 135-145.

Engeström, Y. (1994). The working health center project: Materializing zones of proximal development in a network of organizational learning. In T. Kauppinen & M. Lahtonen (Eds.) Action research in Finland. Helsinki: Ministry of Labour.

Ericsson, K. A., Krampe, R. Th., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychological Review, 100, 363-406.

Freire, P. (1970). Pedagogy of the oppressed. NY: Herder & Herder.

Hatano, G. & Inagaki, K. (1986). Two courses of expertise. In H. Stevenson, H. Azuma, & K. Hakuta (Eds.), Child development and education in Japan (pp. 262-272). New York: W. H. Freeman.

Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-Domain Development of Scientific Reasoning. Cognition and Instruction, 9, 285-327.

Kuhn, T. S. (1970). The Structure of Scientific Revolutions, 2nd ed. Chicago: University of Chicago Press.

Kuhn, T. S. (1957). The Copernican Revolution: Planetary astronomy in the development of western thought. Cambridge, MA: Harvard University Press.

Luchins, A. S. (1942). Mechanization in problem-solving: the effect of Einstellung. Psychological Monographs, 54(6) Whole No. 248.

Martin, T., & Schwartz, D. L. (in press). Physically distributed learning: Adapting and reinterpreting physical environments in the development of the fraction concept. Cognitive Science.

Mead, G. H. (1899). The working hypothesis in social reform. The American Journal of Sociology, 5, 369-371.

Medawar, P. B. (1979). Advice to a Young Scientist. Harper & Row, New York.

Phillips, D. C., & Burbules, N. C. (2000). Postpositivism and Educational Research. Lanham: Rowman & Littlefield Publishers.

Popper, K. R. (1968). The logic of scientific discovery. New York: Harper and Row.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for learning: The hidden efficiency of original student production in statistics instruction. Cognition & Instruction, 22, 129-184.

Schwartz, D. L., Bransford, J. D. (1998) A Time for Telling. Cognition & Instruction, 16, 475-522.

Schwartz, D. L., Bransford, J. D., & Sears, D. A. (2005). Efficiency and innovation in transfer. In J. Mestre (Ed.), Transfer of learning from a modern multidisciplinary perspective (pp. 1-52). Greenwich, CT: Information Age Publishing.

Searle, J. R. (1995). The Construction of Social Reality. New York: Free Press.

Shavelson, R. J., Phillips, D. C., Towne, L., & Feuer, M. J. (2003) On the science of education design studies. Educational Researcher, 32(1), 25-28.

Simon, H. A. (1996) The Sciences of the Artificial, 3rd ed. Cambridge: MIT Press.

Toulmin, S. E. (1972). Human Understanding. Princeton: Princeton University Press.

APPENDIX

Double Transfer Assessment Items (Schwartz & Martin, 2004)

Worked Example Resource Problem Randomly Placed in Half the Posttests

A standardized score helps us compare different things. For example, in a swim meet, Cheryl's best high dive score was an 8.3 and her best low dive was a 6.4. She wants to know if she did better at the high dive or the low dive. To find this out, we can look at the scores of the other divers and calculate a standardized score.

	High Dive	Low Dive
Cheryl	**8.3**	**6.4**
Julie	6.3	7.9
Celina	5.8	8.8
Rose	9	5.1
Sarah	7.2	4.3
Jessica	2.5	2.2
Eva	9.6	9.6
Lisa	8	6.1
Teniqua	7.1	5.3
Aisha	3.2	3.4

To calculate a standardized score, we find the average and the mean deviation of the scores. The average tells us what the typical score is, and the mean deviation tells us how much the scores varied across the divers. Here are the average and mean deviation values:

	High Dive	Low Dive
Average	6.7	5.9
Mean Deviation	1.8	1.9

The formula for finding Cheryl's standardized score is her score minus the average, divided by the mean deviation. We can write:

$$\frac{\text{Cheryl's score} - \text{average score}}{\text{mean deviation}} \quad \text{OR} \quad \frac{X - \text{mean of x}}{\text{mean dev x}}$$

To calculate a standardiazed score for Cheryl's high dive of 8.3 , we plug in the values:

$$\frac{(8.3-6.7)}{1.8} = 0.85$$

Here is the calculation that finds the standardized score for Cheryl's low dive of 6.4.

$$\frac{(6.4-5.9)}{1.9} = 0.26$$

Cheryl did better on the high dive because she got a higher standardized score for the high dive than the low dive.

- Cheryl told Jack about standardized scores. Jack competes in the decathlon. He wants to know if he did better at the high jump or the javelin throw in his last meet. He jumped 2.2 meters high and he threw the javelin 31 meters. For all the athletes at the meet, here are the averages and mean deviations:

	High Jump	Javelin
Average	2.0	25.0
Mean Deviation	0.1	6.0

Calculate standardized scores for Jack's high jump and javelin and decide which he did better at.

Example of a Target Transfer Problem in the Posttest

Susan and Robin are arguing about who did better on their final exam last period. They are in different classes, and they took different tests. Susan got an 88 on Mrs. Protoplasm's biology final exam. In her class, the mean score was a 74 and the average deviation was 12 points. The average deviation indicates how close all the students were to the average. Robin earned an 82 on Mr. Melody's music exam. In that class, the mean score was a 76 and the average deviation was 4 points. Both classes had 100 students. Who do you think scored closer to the top of her class, Susan or Robin? Use math to help back up your opinion.