THE NATIONAL CENTER ON
SCALING UP
EFFECTIVE SCHOOLS

Designing and Scaling Highly Effective Interventions that Produce BIG Improvement:

Counter-intuitive Lessons from the Higher Order Thinking Skills (HOTS) Project

Stanley Pogrow

Conference Paper
October 2015

VANDERBILT
PEABODY COLLEGE

EDC Learning transforms lives.

FLORIDA STATE UNIVERSITY 1851

Georgia State University

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

The National Center on Scaling Up Effective Schools (NCSU) is a national research and development center that focuses on identifying the combination of essential components and the programs, practices, processes and policies that make some high schools in large urban districts particularly effective with low income students, minority students, and English language learners. The Center's goal is to develop, implement, and test new processes that other districts will be able to use to scale up effective practices within the context of their own goals and unique circumstances. Led by Vanderbilt University's Peabody College, our partners include The University of North Carolina at Chapel Hill, Florida State University, the University of Wisconsin-Madison, Georgia State University, the University of California at Riverside, and the Education Development Center.

This paper was presented at NCSU's second national conference, Using Continuous Improvement to Integrating Design, Implementation, and Scale Up. The conference was held on October 7-9, 2015 in Nashville, TN. The author is:

Stanley Pogrow
San Francisco State University

This research was conducted with funding from the Institute of Education Sciences (R305C10023). The opinions expressed in this article are those of the authors and do not necessarily represent the views of the sponsor or the National Center on Scaling Up Effective Schools.

Abstract

There is little discussion in the Design-Based Research (DBR) literature on how to design an intervention that has the potential to be highly effective. The act of designing is usually viewed as engineering something from theory or research on best practices. This paper challenges that universal belief and presents successful design as an intuitive creative process that has little to do with existing academic theory or research—yet is till within the domain of science. Evidence for this perspective is based on (a) the author's experience in designing and disseminating the Higher Order Thinking Skills (HOTS) project which has been one of the most successful large-scale improvement networks, (b) research on the design of the Carnegie Foundation's Statway project, and (c) alternative modes of discovery in science. Implications for the design of more effective interventions and related scholarship are discussed.

Introduction

The focus of this paper is: How do you design an intervention that has the potential to make a BIG improvement in practice across a wide variety of contexts? By BIG I mean that there is an improvement that is clearly discernable and not one that requires sophisticated statistical analysis to tease out. Pogrow (2015) shows how tests of statistical significance or an Effect Size of .2 that is the basis of most conclusions about effective practices generally do not have "practical significance". In other words, these types of small differences/improvements provide no compelling reason why practitioners should switch to using them. The absence of compelling evidence to support existing interventions is a particular problem post-grade 3. For example, the research on intensive reading interventions by Scammacca, et al. (2013) and Wanzek, et al. (2013) found substantial effect sizes for grades k-3 and only non-substantial ones post grade 3. For research to impress practitioners the amount of improvement should meet the "eye" test; i.e., it is obvious to most that a substantial improvement has occurred without needing sophisticated statistical analysis to make the case.

As a result, discussions about how to scale-up innovations need to stop and ask: Is there any evidence that the intervention has produced a BIG improvement? A related question is why should scholars expect practitioners or anyone to change behavior and adopt something that only produces a minor improvement? However, an even more fundamental question is: How does one design an intervention that has the potential to produce BIG improvement?

Alas, there is almost no discussion in the Design-Based Research (DBR) literature on how to come up with the initial design idea. Most of the discussion is about the iterative research process and scaling techniques. But what is the fundamental nature of highly effective education design? The reason there is so little discussion about design is that it is universally assumed that it derives from the engineering of existing academic theory and research evidence of best practices. For example, in their path-beaking work Bryk, Gomez, Grunow, and LeMahieu (2015) describe the iterative research and scaling process of a highly successful application of improvement science to education without indicating how the design idea originated. The belief is that if you have the "right" academic theory or research the design will naturally flow from this knowledge. It is viewed as akin to how someone would design an airplane using the latest principles of aeronautics. Design is thus viewed from the lens of the traditional conception of the scientific process that has dominated educational research to this point. This paper challenges that universal belief and presents successful design as an intuitive creative process that has little to do with theory. It is akin to trying to design the first successful airplane when only very general, largely unsubstantiated, theories exist and available data are incomplete or wrong.

The perspective of the nature of design presented in this paper initially evolved from designing and disseminating the Higher Order Thinking Skills (HOTS) project to reform Title I instruction (which later also included mildly impaired learning disabled students) over the past 30 years. HOTS has been one of the most successful scaling up continuous improvement national education networks in terms of results, policy impact, and number of sites. When I started HOTS I admittedly had no idea what I was doing—and yet it worked far better than what existed and has been subsequently developed. How is that possible? If conventional wisdom is correct HOTS should not have worked. Indeed, while this work presaged the popular emergence of the Design-Based Research (DBR) movement by several decades, there is little that I read about DBR that relates to what we did. Under the predominant conceptions the HOTS design process was wrong. Or was it?

In addition to this author's experience with the HOTS project, this paper draws on research about the design of the Carnegie Foundation's Statway project; another project that has produced BIG improvement at scale. The paper also draws on non-traditional forms of scientific discovery. Together, these sources suggest the need for a different conception of the design process. This paper challenges the predominant conception, or non-conception, about the design process of interventions that end up producing consistent BIG improvement in practice at scale.

Intuitive Design As Science

How do you design the first successful airplane when there is no developed theory and all the existing data about lift are wrong? The answer is that scientific progress has always depended on the use of metaphor, accident, and persistent tinkering. This alternative process of science was what enabled the first successful airplane to be developed by of all things—bicycle repairers.

How do you design an approach to improve the success of Title I students post-grade 3 when all prior efforts had failed? In retrospect, the design of HOTS drew on these alternative processes of scientific invention and knowledge generation. This is what enabled someone who knew little about Title I and reading methods to lead the design of a far more effective approach. (The specifics of the HOTS design process will be described later in this paper.) Unfortunately, our profession does not yet recognize such an approach as scientifically valid. We are locked into a limited model of what constitutes science which limits its how it is applied to improve education.

The Traditional View of How to Apply Science to Educational Improvement and Design

The classic conception in education of how to discover and develop better interventions and new knowledge is to provide a theoretical basis for an approach, implement it, and then test the effects through a rigorous experimental research design that can establish *causation*. Hereafter this approach will be referred to as the education' *"traditional model of scientific discovery and knowledge generation for improving practice"*, or the "traditional model" for short. This traditional model has formed the sole basis of how research, theory, and quantitative methods have been applied to educational improvement—and now to innovation design. This traditional model has predominated because of the perception that this is how science and medicine develops and validates new ideas, tools, and interventions—and education's desire to be perceived as a profession that has scientifically validated practices.

Actual examples of the successful application of theory to solve a fundamental problem are rare in education.[1] There are three main problems/limitations with applying this traditional model of innovation in education.

[1] A notable exception is the successful application of theory for assigning students' to their top high school choices in New York City Public Schools. The approach is based on a theory—but not education theory. The "deferred acceptance" allocation model came from Game Theory in economics. It was developed by three economists to solve the theoretical problem of how to insure that every man and woman gets a preferred choice

1. There is no evidence that interventions based on academic theory work any better than those that are not;

2. The rapid turnover of educational theories with little or no supporting evidence make it difficult to figure out how to decide which theory to apply; and

3. Even if you know which theory to apply, the general nature of most educational theories makes it difficult to establish the specific parameters of use that are likely to be effective out of the almost infinite set of options—so in the end you do not know how to apply it.

For all these reasons Sandoval and Bell (2004) note that it has never been simple to actually translate theoretical insights into educational practice.

Clearly, many seemingly intractable problems remain in education. Researchers tend to blame this on practitioners not using available research on evidence-based practices. However, increasingly some of the most prestigious quantitative researchers in education, as well as in other disciplines such as Bryk, Gomez, Grunow, and LeMahieu (2015) and Berwick (2008) are questioning whether the traditional approaches to applying the scientific process in education is likely to produce such knowledge in the near future. They are not rejecting science or quantitative research. They are simply drawing on other traditions of science and knowledge generation that have not been incorporated to date into education research.

Fortunately, the assumption in education that the traditional model is universally how discovery is made in science is wrong.

Working from academic theory is but one of two approaches used in science. The other approach is characterized by Randall (2005a) as "model building". In addition, many of the important breakthroughs in science have occurred via the alternative paths of: accident, metaphor, and doggedly persistent tinkering. As a result, we should not have to rely on the limited applicability of education's traditional model to harness scientific discovery in the pursuit of better approaches to improving practice.

Alternative Paths to Scientific Discovery

Accident

The classic case of an accidental discovery was Madame Curie discovering x rays because a photographic plate was left uncovered. The discovery was made because unlike most who might have dismissed it as a faulty plate, she was intensely curious as to what might have caused the shadow and was persistent in trying to find an explanation. Many other discoveries such as Teflon were made by accident. "Serendipity is the mother of invention."

Metaphor Based Discovery

in a mass betrothal. The application of this theory and its algorithm dramatically increased the percentage of students who were assigned to a high school of their choosing. This illustrates how a theory for solving a largely theoretical problem can result in an important real-world solution to an education problem.

There are many examples in science where metaphor and intuition, not theory, led to major discoveries. The first manned flying machine was not developed by a theorist applying theoretical principles. The key knowledge that ushered in aviation was discovered by a pair of bicycle builders and repairers. How did they succeed when everyone else before them had failed? Their breakthrough design, the flexible wing came to them by observing birds in flight. They noticed that when birds quickly changed direction they bent their wings. They were then able to use their knowledge of materials and pulleys to design a flexible wing. The wing was based on metaphor…not theory. While the concept of aerodynamics existed, and the general concept of lift was known, it was not really developed. In addition, the Wright brothers' breakthrough came because they discovered that whatever data or predictions existed about the nature of flight were wrong.[2] Only after manned flight was demonstrated did the theory of aerodynamics begin to evolve. In other words, metaphor based invention preceded the development of an elaborated theory.

Manned flight is not the only case where new practice preceded theory. James Watts' invention of the steam engine stimulated the development of thermodynamics. There are numerous cases historically where major inventions emerged from intuitive leaps of the imagination as opposed to the scholarly application of academic theory.

In addition to enabling invention to precede theory, metaphor has also played a role historically in the conceptual development of some of the most important theories in science; what physics calls "thought experiments". The classic example is Einstein imagining what it would be like to travel on a beam of light which helped lead to the theory of relativity.

Iterative Clinical Trials/Tinkering

Many of the major discoveries in science and improvements in other fields were the result of iterative clinical tinkering. The classic case is Edison's invention of the light bulb. It was only his dogged tinkering with various combinations of materials that led to this and his many other discoveries. (James Dyson claims that it took 5,177 tries to develop a bagless vacuum cleaner that worked efficiently.) Most theorists and scientists would argue that what Edison did was very inefficient and certainly not intellectually elegant. That is true. But it turned out to be far more efficient than waiting for science and theory to evolve to the point where it was obvious how to produce a light bulb. How long would the invention of the light bulb been delayed if not for Edison's "tinkering"?

Indeed, it is only when theory has a very strong evidentiary basis does it become more efficient for solving a problem than iterative tinkering. It can also be argued that trying to apply a variety of theories that cannot meet a strong evidentiary standard are themselves little more than iterative tinkering: albeit tinkering with ideas rather than physical models.

Of course, modern medicine would never resort to iterative tinkering. Wrong! Gawande (2007), in his book *Better: A Surgeon's Notes on Performance*, provides a powerful example of how clinical tinkering in medicine has saved lives. He argues that the single greatest improvement in medical practice in the past 50 years in terms of saving lives has been the tremendous reduction in the mortality rate of newborn infants. In the 1950s the mortality rate for newborn infants in the U.S. was 1 in 30. By 2000 only 1 baby out of 500 newborns died. Did this improvement occur from the application of theory? The use of gold standard randomized experiments? The use of evidence-based practices? The answer to these questions is "NO"!

While this life-saving improvement was occurring obstetrics was ranked dead last among all medical specialties in the use of hard evidence from randomized clinical trials. It was considered to be

[2] Once the Wright brothers realized that the existing principles and data were wrong, they built their own wind tunnel and did their own calculations.

the scientific backwater of medical practice. Yet, obstetrics did a better job of improving practice than any other medical specialty. How did obstetrics do it?

The key was the development and widespread adoption of a standard checklist to assess the baby's condition 1 and 5 minutes after birth. This provided an incentive to develop new practices to improve baby's health in the first 5 minutes instead of just leaving the health-challenged babies to die after delivery. How did these new practices develop? Gawande (2007) describes it thus:

> In obstetrics…if a new strategy seemed worth trying, doctors did not wait for research trials to tell them it was all right. They just went ahead and tried it, then looked to see if results improved. Obstetrics went about improving the same way that Toyota and General Electric went about improving on the fly, but always paying attention to the results and trying to better them. Whether all the adjustments and innovations of the obstetrics package are necessary and beneficial may remain unclear…But the package as a whole has made child delivery demonstrably safer…despite the increasing age, obesity…of pregnant mothers. (p. 189-90)

This is essentially crowd sourced clinical tinkering across a network. Different hospitals tried new approaches and communicated what worked in real time to all other obstetricians. How many more babies and mothers would have died if not for the clinical tinkering of hundreds of doctors and nurses?

Recognizing Alternative Pathways of Scientific Discovery in Education

Current conceptions of quantitative methodology and criteria used by funding agencies in education, are universally based on education's traditional model in the mistaken belief that it is the only way that true science can and must be conducted. However, all the previous examples demonstrate that many major improvements in knowledge and invention in science and medicine often do not emerge from the application of theory or from clinical trials/experiments with random assignment. Given that (a) many seemingly intractable problems remain in education, and (b) the historical importance of alternative paths of discovery and knowledge generation in science, it seems necessary and appropriate to have an alternative track for generating new innovative practices and knowledge in education. This does not mean that the traditional model and theory driven research is bad. Only that they should <u>not</u> be the <u>only</u> way, or possibly even the predominant way, that funding and academic prestige is allocated in education.

Despite the existing limited conception of science and knowledge generation in our field, some interventions that have been generated outside the traditional model have started to permeate education. For example, the Kahn Academy was not theory-based. However, this and other widely used non-traditionally designed interventions have emerged largely from outside the formal education academic institutions and funding sources. Indeed, the design approach of the Kahn Academy lessons violates most theories taught in colleges of education about how to apply technology in instruction and how to design effective instructional materials and user interfaces. Instead, Kahn used his intuition as an engineer to develop his tutorial lessons, and the result was a very basic, almost primitive, interface.

Is there a formal way to incorporate the role of accident, intuition, metaphor, and clinical tinkering to the improvement of educational practice? The closest we have come is the emergence of "Design-Based Research".

What is Design-Based Research?

Design-Based Research (DBR) is academicians, researchers, and practitioners coming together to design a novel approach to solve a problem, test the effects of the approach, and use feedback to make iterative improvements. The recommended characteristics for DBR according to Mingfong, Yam San, and Ek Ming (2010) are (a) using mixed methods, (b) multiple iterations of the design, (c) collaborative partnership between researchers and practitioners, and (d) evolution of design principles. The research feedback indicates what is working and what is not, and iterative changes are made with the goal of continuous improvement. (NOTE: "Iterative changes" is a fancy way of saying "tinkering".)

DBR is an iterative collaborative process of learning by doing, i.e., trying out something, and improving it based on repeated experience—as opposed to a single grand controlled experiment. The research goal is to refine the intervention across multiple iterations and increase the number of sites to test its effectiveness and scalability.

The start of DBR is generally credited to Ann Brown (1992) who came to realize that results from laboratory based research were inherently limited in their ability to explain or predict learning and moved her research to the classroom; a process she called "design experimentation". In their history of DBR Sandoval and Bell (2004) note how others in educational psychology, the keepers of the flame for the scientific model for quantitative research, began to question the traditional model of quantitative research. They quote, Lagemann (2002) as noting that, "…the traditional paradigm of psychology has striven for experimental control at the expense of fidelity to learning as it actually occurs. Thus, although such claims might be scientific in one sense, they do not adequately explain or predict the phenomena they purport to address…" (p. 199).

The DBR movement began to pick up steam as it moved from educational psychologists to a broader audience of educational researchers. In 2003 *Educational Researcher*, the journal that goes out to all members of the American Educational Research Association (AERA), had a special issue devoted to DBR (Kelly, 2003). According to Bell (2004) scholars from a wide variety of disciplines became interested in participating in DBR to: "… better understand how to orchestrate innovative learning experiences among children in their everyday educational contexts as well as to simultaneously develop new theoretical insights about the nature of learning" (p. 244). Collins (1992) described educational research as a "design science," like aerospace engineering that required a methodology to systematically test design variants for effectiveness. Sandoval and Bell (2004) introduced the concept of "embodied conjectures". These are conjectures (rather than formal hypotheses as generally used in experimental research), and it is about learning within educational designs. While a preeminent focus remained on implications for theory development, there was increasing recognition of the need for an alternative to the traditional controlled type of learning environment in the form of an approach that took place in authentic settings, relied on quick testing of emerging designs, and that was based on conjecture as opposed to formal hypotheses.

Interest in DBR began to accelerate in the academic community. A literature review by Anderson and Shattuk (2012) found that the number of articles that discussed DBR increased from almost zero in 2000 to almost 400 in 2010, and that after 2006 the nature of the articles shifted from discussing the characteristics of DBR to reporting the results of DBR research. In addition, many who do this type of work are not aware of the term DBR—so the results of this survey are probably an underestimate of the amount of DBR type scholarship.

As this is being written, there seem to be three emerging perspectives on DBR.

1. Traditionalist DBR. "Traditionalist DBR" consists of individuals who are trying to define DBR as simply research that is based in the real world while maintaining all the other elements of the traditional model. For example, they insist on the pre-eminence of theory as

justification for the design of the intervention and on judging the results on the basis of whether it generates new theory. Anderson and Shattuk (2012) and McKenney and Reeves (2013) treat the DBR outcomes of improved intervention and theory generation as of equal importance. The advocates of this perspective are having great influence on the funding criteria currently being used by ED to evaluate DBR funding proposals. The problem is that the traditional model of generating knowledge is too limited in other aspects of meeting Langemen's challenge of doing research that has "…fidelity to learning as it actually occurs" which involves more than simply where the research is conducted.

2. Design Based Implementation Research (DBIR). The focus of DBIR (Fishman, Penuel, Allen, & Cheng, 2013; Russell, Jackson, Krumm, & Frank, 2013) is to collaboratively design interventions that are then used as the basis for understanding the conditions under which practitioners decide to accept or reject new interventions. The goal is to generate new and better implementation theory.

3. Networked Improvement Community (NIC). The goal of an NIC (Bryk, Gomez, & Grunow, 2011) is to design a new approach that produces a BIG improvement in a major education problem. The new design is piloted in one or more initial sites, and as iterative improvements are made to the design, it then gets expanded to other pilot sites, and the iterative changes are communicated to all the sites in the network.[3]

Limitations of DBIR and Traditional DBR

I have great misgivings about the DBIR and Traditionalist DBR approaches. After meeting some of the DBIR leaders at the 2013 AERA meeting I engaged in a 6-month email debate. My key concerns were as follows:

- The interventions being presented as examples of DBIR projects in many cases appeared to be a rehashing of pet ideas and theories by professors that had not worked in the past and which were being reincarnated under the umbrella of *DBIR*;

- Some of the needs that the projects were attempting to address were way too general, e.g., improve teaching;

- The researchers advocated for the traditional attitude that designs should be based on academic theory. They rejected the contention that "design" was an artistic/intuitive process (as most people would think of the term "design"). These researchers saw design as the iterative engineering of theory. I think that Steve Jobs would disagree if he were in a position to do so.

- They were not able to provide any examples where interventions based on theory worked. They justified their conception of the use of theory by citing lots of theory to demonstrate the importance and centrality of theory.

[3] I have taken some liberties with The Carnegie Foundation's technical descriptions of some of the key elements and processes of NICs in order to present this somewhat simplified version of what I think the key essence of their approach is, and to use language that is more familiar to readers. In order to be respectful to the details of Carnegie's conception I use the phrase "NIC-like" in several places.

- However, the BIGGEST problem was in relation to DBIR's intent to study why so many individuals decide not to implement a new intervention. However, they were in denial as to whether it is reasonable to expect practitioners to implement an intervention favored by researchers if it does not produce a BIG improvement: and such denial will necessarily bias DBIR researchers' conclusions about practitioner implementation decisions.[4]

In the end, the failure of the DBIR researchers failure to understand the critical importance of BIG improvements essentially means that they do not have an adequate theory as to the conditions under which it is reasonable to expect leaders and staffs to adopt a new practice. As a result, practitioners will be improperly judged as at fault if they do not adopt the intervention designs that the DBIR researchers deem to be scientifically validated—even if there is only an actual SMALL advantage demonstrated via statistical significance. This will continue the historic problem of academicians judging interventions to be successful based on inadequate statistical criteria and then miscategorizing practitioner's reluctance to implement them as a resistance to applying research findings. This is the equivalent of expecting people to switch to a new car simply because it gets 1-2 more miles per gallon.

As a result, DBIR and Traditionalist DBR appear to essentially be business as usual movements. They use the new rationale/strategy of DBR to once again apply favored theories while maintaining most of the elements of the traditional research model and innovation design process. Unfortunately, these branches of DBR have dominated the establishment of the criteria for determining (a) what DBR is, and (b) how DBR design, research, and development should be funded.

On the other hand, Networked Improvement Communities (NICs) branch appears to offer the greatest potential for DBR to use a flexible approach to design and thereby provide the potential to develop more effective interventions by applying the alternative pathways of scientific discovery and more authentic research methodologies. The next few sections describe the process of designing and researching NICs in more detail.

Origin and Goals of the NIC Concept

The concept of an NIC developed by the Carnegie Foundation evolved mainly from the work being conducted in medicine and health care under the rubric of *Improvement Science*. Research in health services geared to improving patient results began to change in the previous decade. Some in the medical field, such as Berwick (2008) and (Plsek) 1999 began to promote the view that if medicine was to improve the quality of health care it had to view the development of better approaches as requiring research methods that take into account the social processes and context specific mechanisms at work in delivering better patient care across institutions. They successfully argued that the gold-standard research methodology, referred to in the medical literature as *randomized controlled trials* (RCT), was inadequate for such analysis. For example, consider the following statement by Berwick (2008) in the Journal of the American Medical Association:

> Changes in the current approach to evidence in health care would help accelerate the improvement of systems of practice…Educators and medical journals will have to recognize that, by itself, the usual… experimental paradigm is not up to this task. It is possible to rely on other methods without sacrificing rigor. Many assessment

[4] To emphasize the importance of BIG improvement I used the example of the Dvorak keyboard. This keyboard was not adopted even though it offered a 25% improvement over the typing speed of the standard QWERTY keyboard. It simply was not a big enough improvement for everyone to give up the entrenched form of the keyboard.

> techniques developed in engineering and used in quality improvement…have more power to inform about mechanisms and contexts than do RCTs, as do ethnography, anthropology, and other qualitative methods. For these specific applications, these methods are not compromises in learning how to improve; they are superior.

These alternative methods of scientific discovery and assessing evidence of effectiveness are now being used in a variety of networks springing up in medicine and other fields to try and solve heretofore intractable problems. These methods are equally applicable to solving problems in education.

The goal of an NIC is to solve a major existing problem in education. The existence of many unsolved problems in education suggests that it is either impossible to design solutions with high levels of *external validity* that can work consistently across a wide variety of sites, or that we need a new generation of better interventions. The goal of an NIC is the latter—to try and produce an intervention that can be scaled with predictable and consistent improvement. The key methodological goal of an NIC is to demonstrate that an intervention had high levels of replicable success for solving a clearly defined problem.

Applying improvement science research strategies in education means that the focus of the research is not to conduct formal experiments using gold-standard design and statistical methods to eliminate all possible confounding factors, but to engage in a series of iterative trials in a variety of contexts so that the designers can learn by doing and can quickly modify and elaborate the intervention as needed. The most desired form of evidence is consistent replicated levels of BIG improvement across contexts that thereby demonstrates high levels of *external validity*. While it is harder to actually develop an intervention that works consistently across contexts, as opposed to the more limited context of experimental research, the evidence required to show consistency of effects is simple and basic. You simply need to demonstrate how much improvement occurred in each context and the degree of consistency. If you see similar improvement occurring in very different contexts at different points in time there is virtually no chance that the improvement occurred by chance.

NICs are also different than the traditional reform network in education. In the latter, some type of reform is deemed to be effective and scientifically validated, and everyone is pushed to adopt it. A recent example would be the NCTM math reform. Experts are identified who then roam the nation or a given state training everyone in how to do the reform. Those who adopt indeed constitute a network. In the end such networks generally have little or no effect, and it turns out in retrospect that the reform was not sufficiently defined to take into account all the processes involved in trying to make it actually work. Everyone then moves on to the next reform of the moment. However, an NIC does not claim to have experts in solving the problem of practice that is its focus. Its perspective is that the problem has not been solved because no one has yet figured out how to do it. As a result, the formation of an NIC is viewed as establishing a process wherein "…improvers can systematically learn by doing" (Bryk, Gomez, Grunow, & LeMahieu, p. 205). It is a process based on humility about the limits of what we know as opposed to believing we know what works.

Two examples of successful NICs are the HOTS program organized by this author and the Statway initiative of the Carnegie foundation. How were these designed? Were there common elements in the design process? Were they based on the traditional method of scientific discovery?

The Design of HOTS versus Statway

Comparing the design process of HOTS and Statway is instructive since the design processes were separated by 25 years and were completely independent of each other. In addition, the two projects involved very different contexts: HOTS involved at-risk students in grades 4-8, and Statway involved first year community college students enrolled in developmental math. What binds these two interventions is that they both produced BIG improvements with at-risk students by designing a surprisingly novel approach that eliminated a fundamental roadblock to their progressing academically. As a result, if the design approach for these two very different interventions was similar, that would be a powerful validation of the importance and viability of this approach.

Background of the HOTS Project

The Higher Order Thinking Skills (HOTS) project started 30 years ago to reform Title I. Title I is the largest federal program in education and provides supplemental help to children born into poverty to succeed academically The HOTS project was established to improve the drop-off in progress that at-risk students make after the third grade and reverse their subsequent decline in academic performance. While achievement gaps stay stable or decline during the first three grades, by the eighth grade the gaps have rewidened and are large in all measured subject areas despite decades of reform.

The HOTS intervention replaces the content-based remedial approach to supplemental services for Title I and Learning Disabled (LD) students in grades 4-8 with intensive general thinking development activities. It provides a Socratic learning environment combined with technology to create consistent intensive conversations about ideas. This intervention is the only one provided to students; i.e., all the other supplemental "help" services are eliminated. HOTS daily activities are supported by a detailed curriculum and an intensive teacher training process.

From the beginning, HOTS students consistently made three times the growth in reading comprehension and twice the growth in math on standardized and state tests without extra "help" in those content areas—as compared to students receiving extra instruction in those subjects. Approxinately15% of HOTS Title I students made the school's honor roll within the first year. These gains showed up across all cultures and school settings, and were accompanied with major increases in student verbalization (Pogrow, 2004; Pogrow, 2005). The program was subsequently adopted in close to 2600 schools nationally and served close to ½ million Title 1 and LD students.

The surprising success of HOTS also impacted national policy. When the program first started it was technically illegal for schools to use Title I funds to develop thinking skills. The success of the HOTS program led to a reformulation of the Title I law to require the teaching of "advanced Skills".

HOTS started with small grants from the US Department of Education. The key was annual small grants from the now defunct National Diffusion Network which helped create a marketplace for promising innovations. Later on support was provided by the Ford and Edna McConnel Clark foundations. After approximately 10 years HOTS became self-supporting.

Background of the Statway Project

Statway is an NIC effort to improve upon one of the greatest systemic failures in American education. Approximately half the students who enroll at a community college (CC) are placed into a remedial/developmental math course sequence of at least a year which they must pass before they can enroll in CC credit earning courses. In this sequence they are essentially asked to relearn high school mathematics. Bryk, Gomez, and Grunow (2011) cite national statistics that 60-70 percent of these students nationally fail to complete the sequence of developmental math courses and drop out without having had the opportunity to earn any CC credit. *This may be the highest dropout rate in American education.*

The Statway initiative has designed and tested a new developmental math sequence that emphasizes the statistical types of math knowledge that students need in CC coursework and the types of problems they will encounter in a variety of CC majors (Bryk, Gomez, & Grunow, 2011). The Statway curriculum was developed collaboratively by researchers, foundation scholars, and CC administrators and faculty. Results from the small-scale field tests show that the percentage of students earning community college math credits increased from 15% after <u>two years</u> of developmental math to 50% after only a <u>single year</u> of Statway developmental math. This certainly qualifies as a BIG gain.

What is "Design"?

In both of the above cases the existing approaches for improving the success rate of at-risk students were not working. But why were HOTS and Statway unusually effective and produce better solutions to widely recognized problems than what previously existed? The specific element of the interventions' successes that is explored in this paper is the nature of the design process used, and whether there seem to be general design principles that can be more widely applied. However, before discussing the nature of the design used in these two interventions, it is useful to first explore the nature of design in general.

I am not sure how to specifically define "design" other than that it is not an engineering process; though it can have engineering elements. Design also clearly has intuitive, aesthetic, and metaphorical components. It is also clearly a process that is increasingly critical to success of a wide range of products and activities in our society. The success of Apple products is partly based on the sophisticated use of design. But much of the design was based on the instincts and life-experiences of Steve Jobs—not academic theory.

It is interesting that with all the emerging scholarship about DBR there is virtually nothing about the aesthetic, metaphorical nature of design. It is viewed as an engineering process, e.g., as something in which specific procedures and formulas exist that can be applied. Yet there are many examples of collaboration between aesthetics and science even in the physical sciences. For example, engineering can also include aesthetics. However, a better example is probably architecture. There is a very strong aesthetic element in the design of buildings, but then there is also the materials science and engineering to make sure the building will remain standing. One interesting example is the iconic, Sydney Opera House. The design of Jørn Utzon won in a competition based on aesthetics. Once the design was selected the realization set in that there was no way to actually build it under existing knowledge and available technology. Fortunately, the decision-makers remained committed to the aesthetics of the design and waited for the science and engineering to catch up.

Another individual who has successfully combined artistic design and scientific engineering is James Dyson, the inventor of the bagless vacuum cleaner and hand dryer. His inventions have made him one of the richest people in England and earned him a knighthood. His initial background was as a student in the Bryan Shaw School of Art and the Royal College of Art where he studied architecture. Once he had the design skills he then studied engineering. Like Steve Jobs who studied calligraphy, it is the combination of skills in artistic design and scientific engineering that leads to breakthrough inventions. Dyson sees the engineering and design processes as separate but synergistic processes (Science Friday, 2014): which implies that design is a more intuitive, creative process that is separate from engineering—and often also separate from academic theory.

The successful design of more powerful interventions also requires intuitive and creative leaps of thinking. In Dyson's case, the idea for using cyclonic action as the key to eliminating the vacuum bag and filter was happenstance. The idea came to him while walking by a junkyard and observing a huge cyclonic tower and wondered if a smaller version could be made for use in a vacuum cleaner (Science Friday, 2014).

From my own experience I can now say without false modesty that my real, and probably only, talent is that I am a master designer. To me, a master designer is someone who can think of new approaches and instinctively see in his/her head how diverse students will react to a new idea, a new approach to teaching something, and/or a new process for achieving some outcome. I suspect that the following are some key characteristics of a master designer:

- Immature and adventurous—yet organized,
- Gutsy/fearless,
- Tremendously imaginative with an over-developed sense of fantasy,
- Highly skeptical, and
- Insatiably curious about things in general and especially about how kids view life and their place in it.

While it is not clear what exactly "good" education design is, a better approach is to ask the following answerable question: *What, if any, were the common characteristics of the design process used by HOTS and Statway to develop highly effective interventions?*

The Design of HOTS

When I got involved in a collaboration to design what would eventually become HOTS, I had an advantage in that I did not know anything about Title I, cognitive psychology, or theories of teaching reading. I had no idea as to what "best" practice was or what the latest theories of the moment were. I might as well have been a bicycle builder and repairer.

What drove the design at the start was the insight of the practitioners who approached me to work with them. Their instinctual insight was that their Title I students were bright and they realized that the current remedial approaches were not working.

Based on that insight we decided to design an intervention suitable for bright advantaged kids. (This was technically illegal under the then existing Title I regulations.) The first metaphor used to drive design was to consider our Title I students as individuals attending an elite private school. So we set out to design something that mimicked the type of education we felt that such students would receive—even if only for a small part of the school day. That led to the decision to focus on using an intensive Socratic approach. The second metaphor was to think of the HOTS Title 1 intervention as replacing the conversation that students were not getting in the home. This intuition was supported by the classic study by Hart and Risley (1995) that found that there is shockingly little conversation in caring low-income households and that this appeared to have an effect on their children's cognitive development.[5]

Once the decision was made to create a conversation rich environment, the next design questions were: (a) What kinds of conversations should we create, and (b) what types of conversation stimulate cognitive development? Well, the classic conversation in the home was dinner table conversation. Emulating key aspects of dinner table conversation was a logical extension of the second metaphor and it served to guide how we designed key aspects of the conversation, curriculum, and computer use activities. Since dinner table conversation is an ad hoc exploration of what happened in the context of recent life experience, we used experiences on the computer as a metaphor for life experiences. This third metaphor led us to organize discussions that explored their experience on the computer in a manner that parents question and prod their children around the dinner table. That is why, for example, we made the counterintuitive decision not to link the discussions back to the classroom content since parents in the home do not link their dinner table conversations to classroom content. This metaphor also provided some guidance for the design of the Socratic system.

But how could we stimulate cognitive development? That was solved by the fourth metaphor which was to think of the *brain as a muscle*, and brain circuits as muscle fibers. If athletes engage in building muscle via repetitive actions, could we structure the conversations so that students would repeatedly engage in verbalization that mimicked the physical structures of neural networks? I

[5] The reality was that we did know of this study when we made the decision to focus on creating a Socratic environment. That decision was made solely on the basis that we felt that this was an enduring characteristic of elite education.

approached a series of cognitive psychologists and asked them: “Which 3-4 studies or researchers have produced the best evidence (not theory) on how brain networks work for storing and retrieving information?” Note that I did not try to read all the research on brain functioning or survey all theories. I took the small set of resources they provided me and found that the key research evidence was that learning and retention could be enhanced by increasing linkages between and among concepts. So the basis of all the curriculum and conversations in the intervention was to maximize the linkages between all the disparate experiences students had on the computer with each other (via the use of a set of linkage concepts) and with their existing general knowledge base.

While we used research evidence about how the brain stores information we were not applying theory per se. Rather we used experimental research that fit with the chosen metaphor. So it was the metaphor of the “brain as a muscle” that drove the types of research evidence we were seeking.

A final metaphor was used to develop the teacher training. Socratic teaching requires a different way of listening and speaking to students. How to develop these skills? It was clear that we needed training methods to develop different teacher behaviors and communication interaction reflexes—and that simply providing knowledge and theory about the nature of Socratic interaction would not be very effective.

At the time I was living the Los Angeles area and hanging out with some actors. I was fascinated as to how individuals who did not know what day of the week it was could memorize and perform with great nuance scripts that could easily be more than 100 pages. How was this possible? I spent a sabbatical in the theater school at UCLA to learn how teaching and learning occur in the theater. It turned out to be virtually the opposite of how they are done in schools. Actors learn their lines and how to perform them through the process of contextual familiarity gained from repetitive experience.

This led to metaphor #5, which was to adopt the process of how teaching and learning occur in the theater as the basis for designing the HOTS teacher training. The training is a 5-day experience in which each prospective HOTS teacher “teaches” three lessons to the others who play the role of students responding spontaneously to the questions and followup probes put forth by the teacher. It is through such experience and feedback from peers that these truly excellent teachers come to realize that they previously had not instinctively listened to what their students were saying. Then they become open to learning the very different listening and probing methods of the HOTS Socratic system.

In any event, all of the key metaphors led to very specific decisions as to how to structure the intervention. Someone working from a perspective of academic theory, and even research evidence on best practice, would have shaped the intervention differently. Indeed, 10 people looking at the same theory and evidence would probably come to 10 very different sets of decisions on key design elements of the intervention. However, none of them would have come to the critical conclusion that the discussions should not be linked to classroom content…a critical decision that flowed naturally from the home dinner table conversation metaphor. Nor have any of the other efforts developed from the traditional model produced anywhere near the BIG gains after the third grade that HOTS has.

But metaphors by themselves are not sufficient. Lots of supporting details and materials need to be developed to support a Socratic intervention. This takes lots of hard work, persistence, and open-mindedness on the part of all concerned. While I did not know what to expect, the open-mindedness of all the initial collaborators led to an expanding learning community as the number of HOTS sites increased. While I took the lead in the writing of the materials and the initial design of the workshops, the development process quickly developed a mind of its own. No one individual can make the myriad decisions that go into developing the day-by-day materials needed to support an intensive intervention. As teachers began to use the materials they quickly started to make suggestions as to how they could be improved, ranging from small details to different conceptual approaches to a given lesson. Suggestions came in from teachers all over the US. The creativity of their insights, as well as their suggestions based on their experience in teaching the lessons to their students were invariably better than the original conception. After a period of time a new version of the curriculum would be issued as part of the iterative improvement process. Now when I look at the curriculum I have no clue as to which ideas were mine and which came from teacher recommendations.

The same iterative process happened in the design of the intensive Socratic professional learning development. While the outlines of the initial training held up, the Socratic model underwent continuous tinkering to insure that teachers were prepared for all the different types of situations they would encounter in trying to maintain a questioning environment to stimulate student verbalization as opposed to "telling" key ideas. Teachers needed to have the specific tools and probing strategies to deal with all of the types of student responses and challenges they would encounter. Extensive teacher feedback of what they were encountering and how they were dealing with the situation, as well as extensive observations of classroom conversation, were used to make the needed tweaks in the Socratic model and training.

While some would consider this iterative process to be grunt work, it was in fact an amazingly intellectually rewarding process of idea sharing within a true learning community. The continuous flow of data about student outcomes and implementation processes were critical for both improving the consistency of outcomes and increasing scalability to a wider variety of contexts. This iterative development process lasted about 12 years for the curriculum, Socratic model, and training before they were considered pretty much done. This iterative process is essentially the scientific process of "persistent clinical tinkering". Yet, it is also important to note that the intervention produced unexpectedly large gains from the very beginning—and that the tinkering improved the results only at the margin. Most importantly, the ongoing tinkering was critical for expanding the benefits across a wider variety of contexts.

An example of such "expansion tinkering" was what happened when schools in Salt Lake City Utah adopted HOTS. One of the software games involved a simulation of the street shell game where you bet money that you can identify the shell that the pea is under. This is used to enable students to discover the concept of number series and the benefits of the systematic application of a strategy. However, this conflicted with the Mormon opposition to gambling. However, they did not consider it gambling if the students played for points instead of money. We therefore modified the software so that teachers could choose whether the students played the game for make-believe money or points.

The critical processes in the design and improvement of HOTS were the use of metaphor and persistent clinical tinkering. The key design element was in picking the right metaphors initially. That was partly luck; but it was also a result of the intuitions and personal theories of action of the practitioners involved—as well as the openness and creativity of all involved.

Once the initial design was in place and the results of the initial pilots were promising, it then became possible to tap into some research that could elaborate and extend the initial design. Some of the key work was Vygotsky's experimental data that were consistent with what we were seeing with HOTS students. This ex post facto application of research to explain and extend what was being found in the design is not how academicians tend to think about the application of research to improve practice. Indeed, now when I tell cognitive psychologists of the brain as a muscle metaphor they think that is "moronic". I tell them in response that… "It is not moronic but good design". At that point they generally turn away shaking their heads.

The Design of Statway: Similarities and Differences to HOTS

Was the design of HOTS a unique process or was it similar shared to the design of Statway that also produced BIG, consistent improvement? In order to answer this question I interviewed three of the key individuals behind the design of the Statway project at the Carnegie Foundation. Interestingly, while this group has been in the forefront of applying "Improvement Science" to education and documenting many of its characteristics, they have written almost nothing about the rationale for the initial design conception of Statway. It turns out that there were some differences in the design processes used in the two interventions and many similarities. It is the similarities that have the greatest implications for understanding the ideal characteristics of DBR design.

First the design differences. The development of Statway was funded by large external grants. This made it possible to hire lots of different groups to develop key elements of the project. In addition, the development process was sequential. First the curriculum was developed over an extended period of time. Then once the curriculum was in hand the process of designing how to train teachers began and research began to be brought in as to how to organize instruction: the project began to realize the need to rethink professional/staff development. In HOTS a small group designed all aspects of the intervention pretty much simultaneously. Clearly simultaneous curriculum development, teacher training, and feedback is easier to do in an elementary school supplemental program than it is for inserting a year-long course into a community college course sequence. It helped that Statway had a small core of individuals consistently involved in maintaining the original design vision.[6] Finally, Statway was more specific than HOTS in establishing a very ambitious BIG improvement goal right at the start.

However, the similarities were greater than the differences. The 12 biggest similarities were:

1. Both NICs were focused on trying to solve a specific major problem of practice;

2. The initial design approaches were not driven by academic theory. They were driven by metaphor, vision, instinct, and pragmatics based on the beliefs that (a) there was no point in trying to improve the existing failed approach, and (b) it was important to develop a very different approach if there was to be a "short-circuiting of the major problems that students were experiencing";

3. A very different approach was used, and a key element in designing such an approach was to find something that would engage the students;

4. A key element in designing the curriculum was the making of explicit linkages across ideas and problem-solving contexts;

5. All the aspects of the intervention, e.g., curriculum and training, needed to have very clear and detailed specification;

6. There was extensive development and iterative improvement of all aspects of the intervention;

7. There was extensive reliance on basic traditional metrics to constantly measure the degree of improvement that was occurring and extensive field-based observation to identify

[6] It is possible that if Statway had used a more integrated design process, e.g., developing a single curriculum module and have teachers teach it, they would have been able to anticipate some of the problems with staff development and curriculum design earlier and been able to shorten the design and development process.

implementation problems;[7]

8. No one had any idea of whether the interventions would actually work. Once the initial design approach was established the philosophy was "let's try it out and see if it works" and, if the initial results are promising, engage in continuous improvement;
9. There were very promising results right from the very beginning;
10. As the project scaled-up new adaptations became necessary and possible;
11. It was only after the design was in operation that research on student learning was brought to bear on solving problems or for enhancing the effects of the intervention; and[8]
12. As previously mentioned, once it became clear that the designs were in fact producing BIG improvement, there was an effort to understand why it was working. As a result, the NICs not only generated a new, more effective, scalable intervention—they also produced new knowledge.

In terms of item #2 above, most of those interviewed about Statway talked about "vision" as the key driver of the design. There was also a pragmatic basis for some of the design choices. For example, if they were looking for a different context to teach math they could have designed a course around robotics or measuring environmental changes as the alternative approach. The rationales for choosing a statistics course as the alternative approach were:

- The course would be transferable to 4-year college credit,
- The bulk of mathematics needed for community college majors that are applied to the workplace is "statistics",
- Statistics was already in the community college math curriculum so it would be easier to slip into the developmental course sequence and have students receive course credit for it.

There was no reference to metaphor until one of those interviewed was describing the design strategy to get students to buy into the curriculum and learning process, and referred to the approach as "thinking of the brain as a muscle". Given all the grief that I had received for admitting to the use of this metaphor it was great to see someone else independently expressing the same idea. At the same time, it seemed to be an adjunct to the main design component which was mostly referred to as "vision" built around pragmatic considerations.

In terms of item #3 above, the keys to engaging the Statway students were:

- Telling them that what they were learning was "not math" or that it was not "high-school math",
- Making sure that they experienced high levels of success initially, and

[7] In the case of HOTS, the traditional metric was end-of-year test scores which Title I programs are required to report.

[8] Some of the learning research/theory mentioned to incorporate into the staff development/instructional techniques were the importance of productive struggle, explicit connection, and deliberate practice.

- Showing them how what they were learning was fundamental to all their other courses and were critical job skills.[9]

In terms of item #9, it is important to see very promising results in outcomes right from the very beginning since schools or NICs cannot go through 5,127 attempts to get an approach right as Dyson did. Schools are not a vacuum cleaner; not even a bagless one. I suspect that if the initial pilots do not produce major improvements there will not be major benefits down the road regardless of how many design iterations occur. If you do not see strong benefits right from the beginning it probably means that you are simply using the wrong metaphors/intuitions or tackling too general a problem.

Both NICs used very non-traditional approaches and had to overcome political and organizational inertia. I suspect that it is not likely that designs that are variations on a theme that has already been tried and "scientifically validated" are likely to meet the criteria of being able to provide BIG consistent improvements.

A Model of The Design Process for Developing an Intervention that Produces BIG Improvement

Using the similarities between the design experience of HOTS and Statway it is possible to put forth a model of the design process. Figure 1 below summarizes the key characteristics of the non-traditional approach to innovation and discovery inherent in these NIC approaches.

There are two initial phases in the design process. The first is the moment of inception of the initial design idea. This decision seems to rely on the alternative methods of discovery. It is not driven by academic theory. In the case of HOTS inception idea was to treat Title I students as gifted, and in the case of Statway it was to convert developmental math from the traditional algebra to statistics. However, an idea is not an intervention.

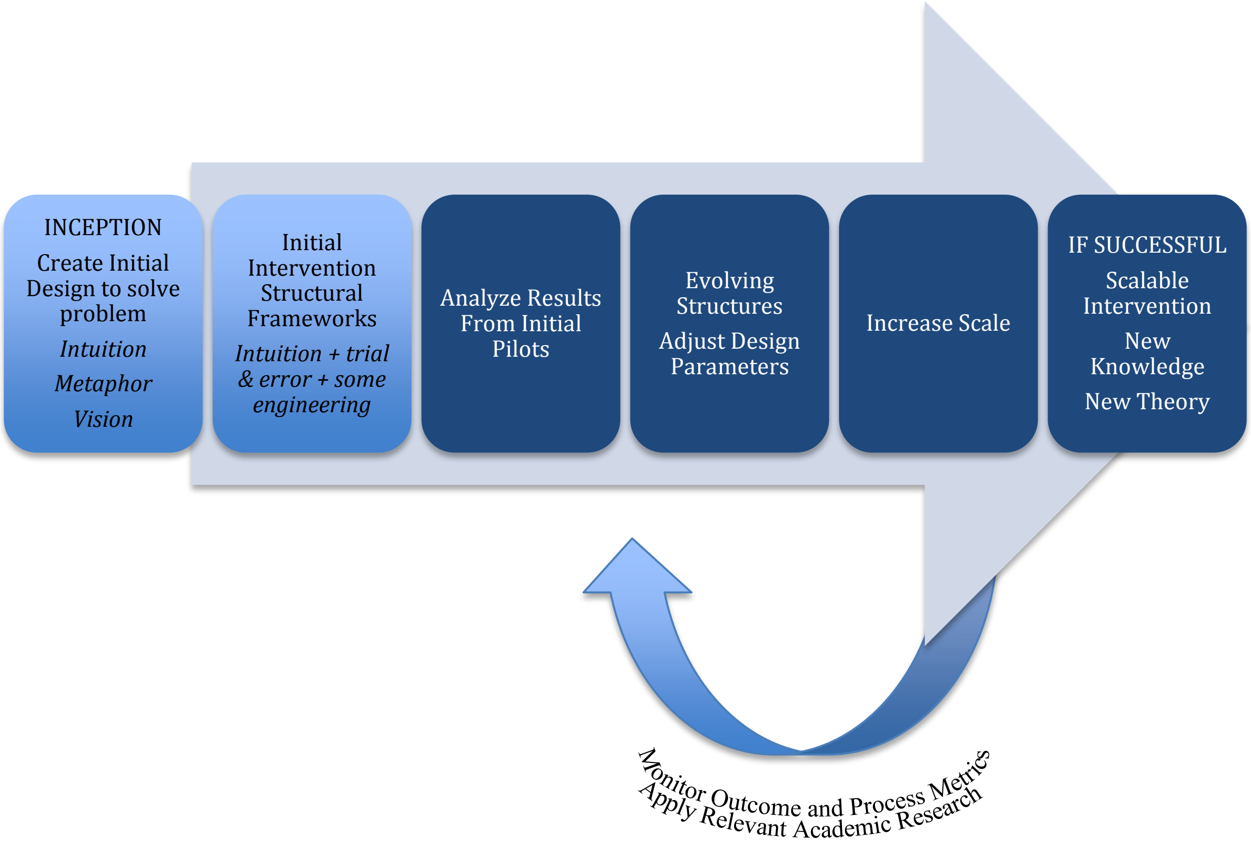
The initial inception idea then has to be fleshed out. Developing the needed curriculum and training involves establishing some initial structural elements in the form of frameworks and working theories of action. In the early stages these frameworks emerge largely from intuition, trial and error, with some formal structures that are engineered. In order to understand these different phases, consider the following two examples:

1. The creation of the universe. The initial inception was the big bang. Then afterwards molecules and particles began to clump together and form primitive clusters. This is the second early design phase of the universe;

2. The previously mentioned Sydney Opera House. The inception was the creative inspiration of the design. Then in the second phase advances in materials science and structural engineering made it possible to design how to actually build it.

[9] The equivalent approach in HOTS for engaging students was to (a) tell students and parents that this was a program for students who the school believes are very bright and is designed to help them bring out that brightness, (b) use technology and provide game-based challenges to students every day, (c) provide opportunity for students to share their ideas and discoveries, (d) have students succeed in mastering a game they initially struggled with, and most importantly (e) provide opportunity for HOTS students to see that their classmates who they always thought were brighter than them cannot solve the problems they succeeded at.

Figure 1
A model of the Design Process


INCEPTION
Create Initial Design to solve problem
Intuition
Metaphor
Vision
Initial Intervention Structural Frameworks
Intuition + trial & error + some engineering
Analyze Results From Initial Pilots
Evolving Structures
Adjust Design Parameters
Increase Scale
IF SUCCESSFUL
Scalable Intervention
New Knowledge
New Theory
Monitor Outcome and Process Metrics
Apply Relevant Academic Research

In the case of HOTS, once the decision was made to actually establish the pilot there was a need to create two-years worth of daily lessons. In order to make this happen there needed to be some organizing principles for the writing process; e.g., focus on making as many linkages as possible across key concepts across the different software contexts. Bryk, Gomez, Grunow, and LeMahieu (2015) referred to these initial frameworks as "primary drivers" (p. 105). These are the starting points of the creation process and the frameworks emerge largely through creative intuition, trial and error, with some engineering of ideas around these frameworks. These frameworks can also be said to be theories of action or working theories. They provide the base from which the supporting materials and processes are engineered. However, even this engineering process requires a great deal of constant creative invention.

However, at this point the frameworks are still largely guesses and hunches. You need to learn quickly from trial and error, with constant feedback, as to what is working and what is not. These feedback loops are critical.

Scaling the Design: Determining the Parameters of Effectiveness Across Contexts

Just because an intervention works in 2-3 initial sites does not mean it will work in 20 or 50 or 500. Once again, it is important to take the humble and honest position of realizing that you have no idea whether it will work anywhere else. The key is to figure out what the specific conditions are under which the intervention will work across contexts; hereafter referred to as determining the *parameters of effectiveness*. Ideally there is a set of parameters that will enable the intervention to be successful in a wide variety of contexts. Most academic theories in education are not specific enough to predict the specific conditions under which the intervention will work consistently. This requires a different form of research process than typically taught in colleges of education. It is a form of learning from iterative experience as opposed to the use of experimental gold-standard research.

Determining the parameters of success

In designing an intervention there are many parameters that can vary. How intensive does the intervention need to be? Which students benefit? Which grade levels or ages? Which types of settings? What kinds of professional learning/development, curricular strategy, technology support, etc. work best. Trying to do experimental research around each of these concerns would be impossibly expensive and take forever. Equally problematic is that traditional experimental research cannot realistically determine which of the infinite possible combinations of the parameters work most effectively together. Nor are leaders seeking proof of *causality*. They need predictive and *external validity*. They need research to tell them the specific conditions, or parameters of implementation, under which an intervention is consistently effective within the uncontrolled environment that they deal with. They can then make an educated guess as to whether it will benefit their school(s).

The problem of relying on randomized controlled trials (RCT), i.e., gold-standard research designs, to inform educational practice is best summarized by Bryk, Gomez, Grunow, and LeMahieu (2015):

> …the on-average difference documented in an RCT tells us nothing about the conditions necessary for better outcomes to occur. Likewise, it tells us nothing about how an intervention can be effectively adapted so that good outcomes can emerge in different settings. This perspective brings different light to common understandings about “what works.” At their best, the rigorous field trials summarized in the What Works Clearinghouse tells us *what can work*… However, these studies provide little or no information about *how to make it work* effectively in the hands of diverse individuals working under varied organizational conditions. Yet this is precisely what practitioners need to know as they seek to implement an intervention in their local schools… (p. 207-208)

So the NIC branch of DBR provides the opportunity to move away from requiring that efforts to improve practice be focused <u>only</u> on experiments conducted with advanced statistics to control for the possibility that a *confounding* variable is causing the observed effects. Having a multiplicity of elements that somehow work together is embraced with no research effort to isolate the relative importance of any individual components or what or why each component contributes to the outcome.

Therefore, in NIC based quantitative research there is little experimental control. Emphasis is on determining the consistency of outcomes, and the extent to which they demonstrate consistent BIG improvements relative to what currently exists across the sites. In an NIC form of DBR the needed implementation parameters for consistent effectiveness are discovered via much simpler forms of research methodology and statistics within the iterative clinical tinkering process. If the intervention works in 2 schools but not a third, why not? What was the difference? How can you adjust the intervention to take this new factor into account? With enough experience the NIC comes to understand the specific parameters and conditions of use under which the intervention is consistently effective. In other words, replicated successful experience replaces the need for most statistical controls. Bryk, Gomez, Grunow, and LeMahieu (2015) refer to this as *practice-based evidence* and note that such improvement research "…is tied to the nature of the problem we are trying to solve—how to get quality outcomes to occur more reliably at scale" (p. 208).

In NIC-like DBR, you use natural variations in the initial phase(s) of implementation to see what is working and what is not working as well. There will be different results and different reactions across the sites. If the intervention is well designed, patterns of effective implementation begin to emerge. These patterns are then incorporated into the next version of the intervention. This process of changing parameters on the fly is the iterative improvement/clinical tinkering phase. It is critical that this iterative improvement process occurs quickly and that the resultant modifications be communicated to all sites using the intervention. Several rounds of iterative improvement will typically be necessary.

An example of how this natural variation in the sites and classrooms leads to better understanding of the conditions of effectiveness is illustrated by what happened when HOTS teachers starting admitting learning disabled (LD) students into the program. (I had advised against this ……… which shows how much I know.) The NIC quickly discovered that it worked with most of the LD students but not all. As a result of teacher surveys and discussions, we quickly came to realize that it provided little benefit to LD students who were below 80 verbal IQ. This meant that (a) it would benefit the vast majority of LD students, and (b) would enable leaders to predict which students needed an alternative intervention. Such a precise finding would have been almost impossible to figure out from experimental research.

An example of a parameter that the HOTS project is now trying to determine whether the intervention will work with students with Asperger's syndrome. Will it work with all such students? Can they be served together with the other at-risk students? Will it be necessary to make changes to components of the program?

We learned how to increase the scale of the program because of the ability to iteratively network a continuous improvement process. It was a growing, participatory dynamic learning community. Everyone in the community was learning from each other. The knowledge gained was then incorporated into subsequent iterations of the program. My role in this networked improvement iterative research process was as an orchestrator who helped create the initial design and general approach and the hub of the information flows from practitioners we were working with all over the US.

The HOTS project used accumulated experience of the emergent patterns of effectiveness to iteratively develop implementation guidelines and standards that were consistently effective across highly diverse settings on a large scale. The iterative process determined with great precision the minimum amount of service students needed in order to develop a sense of understanding that translated to increased academic success as measured by end of year tests and GPA, and to increased social-emotional development.[10] The point is that an iterative process conducted over-time with increasing numbers of sites can produce very precise knowledge about the expected effects and the implementation parameters of effectiveness; something that as a practical matter experimental research cannot provide when there are many complex social interactions.

It turns out that searching for consistent patterns of effectiveness through an iterative process is also used in medicine. This mode of scientific discovery is incorporated into (a) the improvement

[10] Social-emotional development was measured largely by changes in students' verbalization over one –two years in the HOTS environment. This includes their degree of verbal participation, the sophistication and reflectiveness of their verbalizations, and the indicators of self-confidence inherent in how they express their ideas. General experience found that increases in the quality of verbalization were highly correlated with improvements in academic measures. Another measure of social-emotional development was the change in the number of reported friends.

science approach to improving the delivery of medical services, (b) obstetrics, and (c) the frontier of improving cancer rate cures via precision medicine.

Therefore, the best methodological approach for determining parameters of effectiveness where there are many possible combinations and permutations is iterative trial and error without control groups—subject to the condition that the results provide a BIG improvement relative to existing performance levels.

As more experience is gained and more data are received from the initial pilot(s) the frameworks become more formalized and complex, and design parameters are adjusted. The frameworks become structures. As more information is gathered it becomes more apparent what data you need, and the emerging patterns provide a basis for deciding what types of academic research findings are likely to be the most helpful. These findings can then be incorporated into elaborating the design structure. Chances are that the research findings are attached to some academic theory, but the design modifications should be driven by the research findings. So, for example, the research on the importance of the skill of "inference from context" to academic learning caused us to add this as the fourth basic thinking skill in the HOTS framework a year and a half after its inception.

Carrying forward the earlier analogies, at this point in the creation/design process, bits of space particles are now forming stars and planets. A universe is starting to be formed, and the structure of the opera house is starting to go up in Sydney.

At the same time, everything is still dependent on the initial inception idea with no guarantee that this process will actually produce improvement. That makes the results from the initial pilots critical. Lackluster results probably suggest that the initial inception idea has limited potential and that a different approach should be tried. However, if the initial results are positive, then the intervention is spread to new and different contexts in a new iteration, and the materials and parameters are refined. The frameworks become stable structures. Academic research findings are still sought to help make the intervention even more effective. This cycle continues till the intervention has been stabilized. In the HOTS case took it took 10-12 years for each of the key elements to stabilize.

If success continues and the intervention continues to scale, engineering and management issues rise to the fore. More individuals need to be hired to handle fiscal and personnel issues, and increasing components of the project and the intervention structures need to be routinized. The solar system has formed and operas are being performed in the Sydney Opera House.

This multi-phase model of the design process embraces the creative and intuitive parts of science as well as the risky nature of seeking scientific discovery and designing a successful improvement intervention. It parses the parts that are intuitive from those that involve engineering, and from the application of formal academic research to support the quick learning and evaluation process.

Unfortunately, existing descriptions of the design process that focuses only on the formal engineering part make it seem like a rigorous process with a formal specific model. It creates the misperception that if you follow a series of formal steps you are likely to produce a scalable effective intervention. That is a misleading conception. Without the right intuitive conception idea, the intervention simply will not work—no matter how much money is spent or formal procedures get implemented. If HOTS had used different metaphors, or Statway a different initial conception, it is unlikely that they would have been nearly as effective.

Clearly, this design model is based on only two NICs, and it is not intended to mean that it is impossible to design an intervention that produces consistent BIG improvement any other way.[11]

[11] I intend to try and find out the design process for a new, highly promising NIC-like intervention, "Big-

Time will tell.

Some argue that the notion that it is possible to scale-up interventions across contexts is a myth (Berliner & Glass, 2014). That is clearly wrong. A better statement would be using the existing traditional approach to research based on how the medical field tests medicines or that rely on statistical significance is unlikely to produce interventions in education that are powerful and robust. Rather, developing education interventions that scale consistently with BIG benefits requires the use of the methodologies used by the medical profession for improving the delivery of health services—an improvement challenge in an environment with complex social interactions.

In essence what has happened is that the research community has imposed design and research requirements that do not produce powerful interventions that should be scaled. Then when efforts to scale the anemic interventions/reform ideas they do produce aren't effective these same researchers then conclude that there is no such thing as scalable interventions.

At the same time, to be fair to Berliner and Glass, nothing is infinitely scalable—regardless of how effective it is. It is impossible to design an intervention that will always produce BIG gains in all organizations. However, an NIC design that produces BIG gains in an iterative series of pilot studies in a reasonable sample of organizations, and that provides specific guidelines of parameters of effectiveness, will stimulate a self-selection process among potential adopters that will weed out many of the instances where it will not work. Organizations that (a) do not fit the known profile of ones where the intervention was successful, or (b) are not willing to make the changes in cultural perspective and resource allocation that are needed to support the parameters of successful implementation, will choose not to adopt the intervention—if they have the needed knowledge of the parameters of intervention effectiveness.

The HOTS iterative research indicated the specific conditions under which HOTS worked, and these conditions were always communicated in research publications and marketing materials. This enabled leaders to determine if the intervention was suited for their organization and the specific conditions that existed at a given point in time. In many cases, leaders would delay implementation for several years while they worked to make changes to ready their organizations to be able to meet the conditions of effectiveness. Indeed, HOTS would not allow organizations that did not meet the conditions of effectiveness to adopt the program.

The ability to scale an NIC is also a function of (a) whether it provides a substantially better solution to a problem, (b) access to funding, and (c) whether the approach resonates with practitioners and scholars (at that point in time). Policies and beliefs change in education as well as in all other aspects of life. Statway is currently on an upward trajectory of sites but it is too early to tell how large-scale it will become. The small class size movement captured the imagination of policymakers, and a number of states such as California provided extensive supplemental funds to reduce class size. However, the movement petered out due to (a) the high cost, (b) the draining of experienced teachers from the inner city, and (c) reductions in state funding for k-12 education due to the recession of the early 2000s. While the scaling-up process for HOTS was successful for about 23 years changes in federal policy caused the program to lose the vast majority of its sites. The problem was a combination of NCLB and converting Title I into a schoolwide model which caused the Title I community and its district leadership role to dissipate. Does the common core movement provide the opportunity to rescale HOTS? Time will tell!

Implications of This Model of Design for Knowledge Generation and Theory

This description of the design process should not be taken to mean that theory and knowledge

History" to see how it overlaps with the above model of design. This project, which is being funded by Bill Gates, is converting a multi-disciplinary approach to teaching history at the college level to high schools. The goal is to improve the teaching and learning of history.

generation are not important goals of such projects. Once one overcomes the initial shock that something based on creative intuition actually works, and the hard work of starting to scale its use are in process, it is important to understand "why" it is working. The discovery that reducing the content specific skills training in favor of thinking development post grade 3 produced substantially more improvement on content tests is clearly a counter-intuitive finding. I was frankly surprised at first by how well the HOTS students performed on standardized tests. However, this was nothing compared to my astonishment to discover that everything we were doing was the opposite of what the research communities and relevant professional associations think the appropriate way to increase the thinking and problem-solving skills of at-risk students is. Indeed, the approach that was working powerfully 30 years ago is still the opposite of what is universally believed to be best practice for developing the thinking skills of children born into poverty.

The universal conclusion is that thinking and problem-solving must be developed in the context of learning specific subject content: e.g., in the regular math and science classes/periods. For example, Willingham (2007) reviews the research on thinking development and concludes that general/critical thinking cannot be developed independent of the learning of content**.** However, it turns out that this research had been originally conducted with university students who were already highly accomplished in learning the content of their major. What did that have to do with the a 4th grader in Harlem or Appalachia who is 2 years behind in reading and math? Nothing. The other category of research cited by Willingham is very short duration lab experiments of several days to a week with preppy type students. Of course, if you are trying to teach a single concept most efficiently you should do it in content as opposed to teaching some general thinking strategies. This is akin to concluding that after 3 days of laboratory research that there is no benefit from mothers talking to their infants. Some developmental phenomena such as developing speech and patterns of thinking unfold over longer periods of time. The conclusion about how to stimulate thinking development is clearly a case of laboratory-based research misleading the field. It reaffirms the concern that formed the basis of DBR that such research does not capture the nature of learning as it occurs in schools.

Fortunately, I did not know about this research on general thinking. It quickly became clear that the intensive general thinking approach of HOTS was working at a very high level. There are three lessons from this experience:

1. No matter how widely accepted a theory or research finding is in academic circles it may be wrong. In addition, a wrong theory inhibits the development of alternative approaches that may be more effective than what has previously been tried;

2. An intervention designed on the basis of intuition, metaphor, personal theory of action, and feedback, is likely to work better than one based on theory or research on "effective practices"; and

3. If the resultant intervention works it will often lead to the development of new theory—which will challenge existing theory.

In terms of the latter, the success of the HOTS intervention led to the following three theoretical insights:

1. The main reason that at-risk students stop progressing after the third grade is that the curriculum is becoming more cognitively demanding. At that point the students' main impediment is that "they do not understand what it means to understand"—though they have the intellectual potential to do so;

2. A sense of understanding is developed through intensive, mediated Socratic conversation with the critical element being students' active, spontaneous verbalizing of increasingly sophisticated ideas; and

3. *Theory of cognitive underpinnings*. In the absence of a sense of understanding students cannot benefit from quality content instruction to their full intellectual potential or to the extent that students who do have a sense of understanding can. This gap in who benefits from quality content instruction exists regardless of whether the quality instruction is to develop specific content skills or problem solving skills—in all content areas.

Clearly, these theories are controversial. At the same time the *Theory of Cognitive Underpinnings* has major potential for explaining the limited success of progressive reforms to reduce the achievement gap, and also provides a basis for increasing their effectiveness. This theory also calls into question whether the common core will reduce disparities, of be just another example where a progressive reform that focused on developing thinking strictly within content failed to accelerate the learning of at-risk students or reduce the achievement gap.

There are also new insights and surprising results from working at scale and trying to find and understand parameters of effectiveness. For example, I thought that no intervention, including HOTS, would work cross-culture. This turned out to be wrong. We saw the same student reactions in Soldatna Alaska, a small isolated bush village as in inner city Detroit. To me this finding that an intervention could be so widely applicable was an exciting one. Alas, when I included this finding in a recent proposal to the National Science Foundation, rather than recognizing the importance of such a finding the reviewers accused me of being culturally insensitive and rejected the proposal. (It's a good thing that I did not add in that we saw the same results in barrio and Navajo reservation schools.)

A new key insight generated from the Statway project was a better understanding of why community college students who are smart struggle so mightily with mathematics. Givvin, Stigler, and Thompson (2011) documented that the students suffered from "Conceptual Atrophy". Stigler**,** Givvin, and Thompson (2010) define conceptual atrophy as "the willingness to bring reason to bear on mathematical problems lies dormant" (p. 15). This dormancy is viewed by the researchers as a result of prior mathematics instruction that previously failed to connect the intuitive sense that students have about mathematics to mathematical notation and procedures. The students have been conditioned to think of math as the application of rules.

It therefore appears that that when a very novel approach produces BIG improvement in a heretofore intractable problem, new insights results; ones that can contradict widely held belief in the academic community.

Implications of this Model of Design for R & D and Colleges of Education

The good news is that so much of what we think we know from the traditional approach to knowledge generation and personal theories of action are wrong or very incomplete. This means there is lots of opportunity for skeptical, creative people to develop better interventions and designs—which is actually the true fundamental basis of science. However, the design and research processes needed to develop more effective interventions require changes in how education conducts R & D and modifications to the curricula in colleges of education.

R & D

Scholarship and proposals for funding tend to be judged by editors and funders using the existing criteria of education's traditional research model. This is even true for the new DBR funding initiative in the Institute of Education Sciences (IES) in the US Department of Education. Judging DBR (and all other) proposals on the basis of the quality of its theoretical justification gives the process the patina of a scientific rigor and provides a seemingly objective rationale for deciding which proposals to fund. In terms of methodology, top research journals only accept quantitative research that uses increasingly sophisticated mathematical modeling. However, I do not see any new highly effective interventions that have resulted from current education government funding efforts or from such published research. Nor are they resulting in robust forms of new knowledge. It is all tinkering at the margins of knowledge with little or no coherent progress.

The predominant view of scientific research currently used in education has been put forth by education psychology. Design Based Research (DBR) has been a reaction against some of the shackles that that education's traditional approach imposed educational psychology; primarily the locus of research. However, there are other shackles in the traditional view of scientific invention whose value for improving practice also need to be questioned. One of those is that invention design and research need to be driven by theory. While these are important precepts in science, it is incorrect to say that this universally how science is conducted. Of the approaches to DBR, the Networked Improvement Community (NIC) is the most open to removing some of these additional shackles as compared to DBIR because the former takes its cues from the "improvement science" movement in health services as opposed to the methods of testing new drugs.

However, funders and journal editors (along with the vast majority of the research community) have not formalized the importance of intuition, metaphor, dogged clinical tinkering, or knowledge developed accidentally as legitimate data, or a legitimate basis for obtaining funding. Instead, any proposal better have a strong theoretical rationale and it usually needs to be one using the favored academic theories of the moment. Paradoxically, even conceptions of the role of theory in education have been coopted in the service of a too limited conception of science.

My experience is that educational theorists tend not to consider theories produced from NIC-like DBR experience as "real" theories or "real" science. Periodically a theorist will suggest that I put HOTS on a more theoretical basis by doing research to determine how much each component contributes to the overall success of the program. Once again, this shows how the primary compulsions of well-intentioned theorists can distort the application of research to the improvement of practice. First of all, conducting such research would probably cost the entire discretionary budget of ED. Second, why would anyone want to do that? Gavriel Solomon, formerly of Haifa University, would respond to this theoretical excess by pointing out that no one tries to determine how much the string section contributes to the audience's enjoyment of a symphony concert. Why do education theorists think this way?

Education theorists also tend to dismiss findings such as establishing the parameters of how much student verbalization is needed to produce a sense of understanding in the HOTS environment, as a trivial bit of practice based information that has no scientific value. Fortunately, the noted physics theorist Lisa Randall (2005) would disagree. She has defined theory, as a "definite set of elements and principles with rules and equations for predicting how those elements interact. When I speak about theories …I'll be using the word in that sense; I won't mean "rough speculation", as in more colloquial usage" (p. 66). If you substitute the idea of "predicting the interaction of intervention parameters" in her definition, you come to the conclusion that being able to describe precisely that it takes an average of a year and a half of 35 minutes a day of Socratic interaction experience in grades 4-8 to develop a sense of understanding in a specific yet wide range of at-risk categories is in fact a theory in the purest tradition of science—plus it improves practice and helps kids. Conversely, it would appear that much (but not all) of what passes for theory in education falls into the category of "rough speculation".

Indeed, it is possible that in the absence of intuitive master designers involved in developing NICs we may not make progress in the development of more powerful scientific theories in education and will continue to settle for constantly changing "rough speculations". However, iterative research on the establishment of the necessary parameters for an intervention that shows early promise is a particularly critical form of research and theory generation.

However, suppose against all odds you actually succeed in developing a successful NIC, and the series of pilot studies support consistent BIG improvement…what then? Alas, such research will not get published today in any of the top education research journals—regardless of how good the results are. Such journals have made progress in becoming more inclusive in terms of publishing qualitative research. There is now clearly a need for top education journals to expand their current criteria to include quantitative research that does not rely on theoretical justification or advanced statistics and myriad statistical tables to try and establish *causation* or complex relationships—if there are indications of BIG replicated gains.

It is important to finally recognize that in an applied field such as education, producing consistent BIG improvement is an important discovery worthy of publication in the best research journals. It is time for the gatekeepers of quality in education research to come to grips with what medicine and health services has discovered, and that is: Randomized control experimental research design is <u>not</u> the best quantitative approach for understanding how to improve processes at scale that involve complex social interactions with variation across contexts. Rather, the alternative methods of scientific discovery using simpler quantitative methods directly related to measuring the consistency of improvement across iterative trials is the preferred methodology for trying to develop better approaches to solving problems of practice.

Of course, a primary focus on demonstrating *external validity* via consistent gains does not establish *causation*, i.e., that the intervention *caused* the outcomes. The response to that point is: "So what?" While it is heretical to say this, the demonstration of consistent BIG gains across a series of studies in diverse settings is a far more important outcome for leadership decision-making than anything produced by a single gold-standard study.

So accepting the intuitive nature of design within the NIC type framework as a legitimate approach to school improvement and research, and recognizing its unique characteristics, will require non-traditional funding and publishing criteria. This leads to a series of questions: How do you figure out whose designs have the greatest potential and should therefore be supported? If in fact design is an artistic, intuitive, or to use Sandoval and Bell's (2004) term "embodied conjectures" process, how do you determine whose intuitions or embodied conjectures should be funded or published? Which metaphor is worthy of funding? Can art and science co-exist in education research? Once again there are no answers to these questions because they have not been considered.

However, as things currently stand HOTS would not be funded today. Nor can any of my less formal though large-scale improvement research get published in any of the top research journals. Furthermore, practitioner journals have recently been rejecting my work on how to apply the lessons from the HOTS program to improve the success of students born into poverty in the common core. Nor will IES or the National Science Foundation fund my efforts to demonstrate the Theory of Cognitive Underpinnings because everyone knows that general thinking does not work. Case closed!

However, my intent is not to whine. Rather, it is to use my experience to argue that our profession needs a better R & D design mechanism than the standard one for designing and scaling highly effective interventions. The reality is that in my case ignorance of conventional research conclusions was essential. If ignorance of one of the most highly evolved research literature and its overwhelming "evidence" was necessary, what about the utility/validity of all the other less developed research literatures. Because I was ignorant HOTS is the only systematic effort to use intensive general thinking with Title I and LD students that has ever been tested.[12]

How many other creative approaches have not been tried because of incorrect knowledge generated from a too limited conception of science?

It is difficult/impossible to design an intervention to solve a problem that involves complex social interactions using the types of methodologies that are demanded by the top education journals and funding agencies. And even if you somehow did manage to develop an approach that consistently produced BIG improvement it would not be certified by the What Works Clearinghouse in the US Department of education due to its use of traditional research criteria. You cannot design education interventions that produce consistent BIG improvements to an education problem using gold-standard randomized control trials (RCT)—anymore than you can use such methodology in the branch of medicine dealing with the improvement of health services.

The best way to analyze outcomes of an NIC is to simply see how the experimental students did on an absolute basis relative to an existing benchmark. This benchmark can be statewide, national, or even historic results locally. So the question becomes: Does the actual unweighted performance of the experimental students represent a BIG improvement compared to an existing performance benchmark? Even the FDA is now approving treatments for cancer based on such findings without clinical controlled experiments—if the results represent a BIG improvement in cure rates (Kolata, 2015).

Of course, that sets a very high bar for deeming an NIC to be successful: i.e., it is not easy to design interventions that produce BIG gains across sites. However, this may be a far more appropriate bar to set for defining “best” or “highly effective” practices than what currently exists.

I certainly do not know whether the design parameters of HOTS have been optimized, or whether it is even the best way to develop a sense of understanding in children born into poverty. Hopefully, a better designer will come along and develop an even better intervention. However, optimization is not a realistic goal of any complex intervention—regardless of whether it is education or obstetrics. Gawande (2007) noted that while obstetrics has developed their grab bag of clinical practices that have an excellent track record, they do not pretend to have optimized care or to know what the contribution of each component is. They just know that in combination the components with the tried and true parameters of use, their grab bag, reliably saves lives across all the hospitals in the U.S.—and that is good practice.

So the theories generated from NIC, even ones that emerge from an intuitive design process, are valid scientific theories that deserve equal consideration for being formally tested as theories that arise from more traditional scientific processes. When Einstein used metaphor as a basis for developing his theory of relativity, physics recognized the importance of testing the results. Similarly, it is time for education to take seriously the work and resultant theories of intuitive DBR design when it leads to surprisingly successful results. There is also a need to fund the initial testing of intuitive designs, the scaling up of promising ones, and then taking the emergent theories and knowledge as seriously as any other.

However, the way things stand now, the application of the traditional model of science as

[12] The only other research that I know of to even come close is the work of Phillip Adey (1989). This study found that sixth and seventh graders who had 25% of their science instruction devoted to general thinking development over a two year period scored significantly higher on the science achievement test than students who were taught science 100% of the time. While this is only a moderately intensive intervention, it does suggest the potential of general thinking activities to increase science scores.

applied in education is stifling innovation and decreasing the likelihood of developing better solutions to long-standing educational problems.

Colleges of Education

At first glance, many of the important innovations in education such as the Kahn academy are being developed outside of the profession. While our profession is providing ideas such as the common core, it does not appear to be developing powerful innovations. There is a need for colleges of education to make the same type of adaptation that business schools made to train their students to design new approaches in the form of entrepreneurship courses and programs. The equivalent for colleges of education would be to provide training and opportunities in the design of educational innovations to solve problems of practice.

Can “design” skill be taught or even defined more precisely? If it can be taught, what is the best way to do it? How can it be incorporated into the curriculum of colleges of education? Should there be a department of design, and if so what would such a department look like? What kinds of faculty can best teach “educational design”? How do you train individuals to become master designers? What would a design research center in a college of education look like? Indeed, what would a college of educational design look like?

There are no clear answers to these questions as of yet largely because such questions have not been considered in the past. My own experience and reports by other successful redesign efforts suggest that state-of-the-art breakthrough interventions in education will require small teams consisting of (a) experts in academic research who understand the limitations of existing theories and research criteria for determining the effectiveness of interventions (see Pogrow 2015 for a listing of such criteria), (b) individuals who have an intuitive sense of how students and teachers think and react to novel approaches, (c) individuals who can think creatively, and (d) others who can create artistic works. The NIC branch of Design-Based Research provides the opportunity to rethink what a college of education is. There is opportunity to create design centers within existing colleges of education or even to create a new design college of education.

The goal of design centers could be to tackle a major problem identified by a consortium of schools/districts and start from scratch to try and design a new approach/intervention for addressing it. Rather than using existing research and theory as a starting point, it should be a blank sheet design primarily based initially on intuition and metaphor. It should be a collection of highly creative individuals with a variety of skills and knowledge bases, drawn from a variety of disciplines who are interested in tackling the problem. There should also be participation by practitioners. The goal would be to create a very different approach than has ever been attempted previously. Several such design centers could attack the same problem by developing different approaches. While none of the efforts might end up being successful, much will be learned from the failures. In addition, if one of the approaches in fact works, it will be a major breakthrough that probably would not have occurred otherwise.

The development of design centers is not intended to replace the traditional model of scholarship, research, or college of education. It is intended to be a parallel effort. While I currently do not know of any design center in a college of education, there is one at Stanford University: the d.Hasso Plattner Institute of Design at Stanford University. According to its website:

> The d.school is a hub for innovators at Stanford. Students and faculty in engineering, medicine, business, law, the humanities, sciences, and education find their way here to take on the world’s messy problems together. Human values are at the heart of our collaborative approach. We focus on creating spectacularly transformative learning experiences. Along the way, our students develop a process for producing creative

> solutions to even the most complex challenges they tackle. This is the core of what we do.

This is the type of design and invention spirit that is desperately needed in education and colleges of education.

Conclusion

The success of the HOTS and Statway interventions demonstrates that it is indeed possible to design and scale a highly progressive learning environment and maintain BIG improvement outcomes consistently across cultures and highly diverse settings. The experience of HOTS also shows that it is possible to simultaneously produce BIG gains in both traditional test scores and progressive social-emotional development.

This paper has shown that successful design for making major inroads into a key problem of practice appears to be a two-stage process in which the initial stage, the inception, is (a) a highly intuitive insight(s), and (b) that the success of the subsequent elements of design and development are dependent on the viability of the initial insight(s). This model of design calls into questions reliance on education's traditional approach to scientific discovery via the application of a theory or prior research. It also shows how this model of design suggests the need for a different approach to R & D and the generation of knowledge in top research publications than what currently exists,. The paper has also shown how contrary to popular belief, critical branches of medicine have already made such changes.

The medical field has come to recognize that designing interventions to improve the delivery of health services that involve complex social interactions requires the use of very different design strategies and research methods than those used for testing medicines. Improvement science for the former has moved away from reliance on randomized control trials to a gold-standard of demonstrating BIG gains via the use of simpler statistics and research designs that have high levels of external validity. It is time for education to do the same. Indeed, the model of design that emerged in this paper suggests that education's sole reliance on the traditional model of scientific knowledge generation is stymying needed innovation and producing "scientifically validated" practices that do not produce the types of BIG improvements that warrant their large-scale use.

We still need new and better approaches for solving many of the persistent problems in education, and the master designers who can produce them. All we have to do is figure out how to identify and unleash them. Doing so requires that education recognize that new important educational interventions will most likely emerge from intuitive, creative flights of imagination that are based in the alternative modes of scientific discovery. The experience of the HOTS and Statway projects have demonstrated that the alternative modes of scientific discovery and design of metaphor, accident, and persistent iterative clinical tinkering are important methods for generating more effective interventions for solving heretofore intractable problems in education. It is also important for the top education research journals to recognize and seriously consider the different types of research conclusions and theories that emerge from such work.

Proposed new designs should <u>not</u> be funded on the basis whether they are theory based or whether the developer is widely published. Rather, new designs should be considered on the basis of whether something very new is being proposed to try and solve an important problem of practice. This paper also suggests additional criteria for anticipating whether a new design has the potential to produce BIG gains. At the very least we need a federal program to provide small amounts of seed money for such new designs. Such a program should be similar to the now defunct National Diffusion Network program that made the launch of HOTS possible. This program would also be similar to the function of accelerator organizations in business that support promising entrepreneurship ideas with space and small grants to help them get to the next phase.

I want the next generation of designers to have the same opportunity I had 30 years ago. Who knows what they will turn up! In addition, I would also selfishly like to have a shot at getting some of my new designs, such as my Hi-perform elementary school model and my math software for solving math word problems, funded. We have only scratched the surface of the types of imaginative interventions that can be designed to improve education.

References

Adey, P. (1989). Cognitive acceleration through science education." Paper presented to the Learning to Think—Thinking to Learn conference sponsored by the Organization for Economic Development, Paris, July 11-13.

Anderson, T., & Shattuk, J. (2012). Design-Based Research: A decade of progress in education research? *Educational Researcher*, 41(1), 16-25.

Bell, P. (2004). On the theoretical breadth of design-based research in education. *Educational Psychologist*, 39(4), 243–253.

Berwick D.M. (2008). The science of improvement. *JAMA*, 299(10), 1182-1184.

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

Bryk, A.S., Gomez, L.M., Grunow, A., & LeMahieu, P.G. (2015). *Learning to improve: How Americas schools can get better at getting better.* Cambridge, MA: Harvard Education Press.

Collins, A. (1992). Toward a design science of education. In E. Scanlon and T. O'Shea (Eds.), *New Directions In Educational Technology* (15–22). New York: Springer-Verlag.

Fishman, B., Penuel, W.R., Allen, A., & Cheng, B.H. (2013). Design-based implementation research: Theories, methods, and exemplars. National Society for the Study of Education Yearbook (vol. 2), New York, NY: Teachers College Press.

Gawande, A. (2007). *Better: A Surgeon's Notes on Performance*. Metropolitan Books; NY.

Gawande, A. (2009). The cost conundrum: What a Texas town can teach us about health care.

Givvin, K. B., Stigler, J. W., & Thompson, B. J. (2011). What community college developmental mathematics students understand about mathematics, Part II: The interviews. *The MathAMATYC Educator*, *2*, 4-16.

Hart, B. & Risley, T.R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes Publishing.

Kelly, A. E. (Ed.). (2003). Special issue on the role of design in educational research [Special issue]. *Educational Researcher*, 32(1).

Kolata, G. (2015). A faster way to try many drugs on many cancers. *New York Times*, Downloaded 5/12/2015 from: http://www.nytimes.com/2015/02/26/health/fast-track-attacks-on-cancer-accelerate-hopes.html?_r=0.

Lagemann, E. C. (2002). *An elusive science: The troubling history of education research*. Chicago: University of Chicago Press.

McKenney, S., & Reeves, T.C. (2013). Systematic review of design-based research progress: Is a little knowledge a dangerous thing? *Educational Researcher*. 42(2), 97-100.

Mingfong, J., Yam San, C., & Ek Ming, T. (2010). Unpacking the design process in design-based research. In *Proceedings of the 9th International Conference of the Learning Sciences* (Vol. 2).

Plsek, P.E. (1999). Quality improvement methods in clinical medicine. *Pediatrics,* 203 -214.Randall, L. (2005a). *Warped Passages: Unraveling the Mysteries of the Universe's Hidden Dimensions*. Harper Collins: NY.

Pogrow, S. (2004). The missing element in reducing the gap: Eliminating the 'Blank Stare'. *Teachers College Record*, Feature Article, (www.tcrecord.org/Content.asp? ContentID=11381).

Pogrow, S. (2005). HOTS Revisited: A thinking development approach to reducing the learning gap after grade 3. *Phi Delta Kappan*, September, pp. 64-75.

Pogrow, S. (2015). *Authentic Quantitative Analysis for Education Leadership Decision-Making and EdD Dissertations: A Practical, Intuitive, and Intelligible Approach.* National Council of Professors of Educational Administration.

Randall, L. (2005b). Dangling Particles. *New York Times*, OP-ED, September 18. Downloaded 7/27/14 from http://www.nytimes.com/2005/09/18/opinion/18randall.html?pagewanted=all&module=Search&mabReward=relbias%3Aw%2C%7B%221%22%3A%22RI%3A5%22%7D

Russell, J.L., Jackson, K., Krumm, A.E., & Frank, K.A. (2013). Theory and research methodologies for design-based implementation research: Examples from four cases. *National Society for the Study of Education Yearbook*, 112(2), 157-191.

Sandoval, W. A., & Bell, P. (2004). Design-based research methods for studying learning in context: introduction. *Educational Psychologist*, 39(4), 199–201.

Scammacca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2013). A meta-analysis of interventions for struggling readers in grades 4–12: 1980–2011. *Journal of learning disabilities*, 0022219413504995.

Science Friday (2014). Downloaded from http://sciencefriday.com/segment/01/24/2014/james-dyson-failures-are-interesting.html, 10/12/14.

Stigler, J.W., Givvin, K.B., & Thompson, B.J. (2010). What community college developmental mathematics students understand about mathematics. *MathAMATYC Educator*, 1, 4-18.

Wanzek, J., Vaughn, S., Scammacca, N. K., Metz, K., Murray, C. S., Roberts, G., & Danielson, L. (2013). Extensive reading interventions for students with reading difficulties after grade 3. *Review of Educational Research*. 163-195.

Willingham, D.T. (2007). Critical thinking: Why is it so hard to teach. *American Educator*, Summer, 31, 8-19.

Willingham, D.T. (2007). Critical Thinking: Why Is It So Hard To Teach. *American Educator*, Summer, p. 8-19.