The Turing Test:
A New Approach to
Evaluating Investments
in Educational Capacity and Infrastructure

Assessing the Impact of the
Exploratorium's Institute for Inquiry

Dr. Mark St. John

Inverness Research Associates

February, 2001

TABLE OF CONTENTS

The Turing Test:
A New Approach to Evaluating Investments
in Educational Capacity and Infrastructure

Assessing the Impact of the Exploratorium's Institute for Inquiry

Abstract

Inverness Research Associates — a small educational research group located in Northern California — used an innovative evaluation methodology to conduct a rigorous evaluation of the Exploratorium Institute for Inquiry (IFI). The new approach stems from the thinking of the British mathematician, Alan Turing and the acclaimed "Turing Test" which he devised to determine how well computers model human intelligence. Similar to that test, the approach Inverness Research used in evaluating IFI centers on the notion of "distinguishability." More specifically, the approach poses the following research question: To what extent are the participants in a program distinguishable from otherwise similar people who do not participate in a program? More specifically in this case, the question is: Are the elementary science reform projects involved with IFI distinguishable from other, otherwise equivalent, projects that do not have a relationship with IFI? To answer the question in a rigorous manner, Inverness Research set up a "double-blind" study in which both researchers and subjects (in this case the PIs of the NSF funded projects) were not privy to the specific purposes of the research. Instead, a skilled interviewer was given the assignment of interviewing project leaders about the capacity of each project to initiate and sustain a process of inquiry-based elementary science reform. Then independent expert reviewers were asked to review the interviews and make comparative "blind" judgments based on their review of the evidence gathered.

Comparison of the reviewer's ratings show a clear pattern of increased capacities in the projects that were served by the Exploratorium's Institute for Inquiry. Particularly in the depth of their understanding of inquiry, and in the sophistication of their professional development design, the IFI project leaders were clearly and statistically distinguishable from their non-IFI counterparts. The Turing Test in this case clearly provides a way to assess in a rigorous fashion the ways in which and the degree to which the Institute for Inquiry increased the capacity of elementary science reform efforts to do their local work effectively.

Introduction

Overview

This monograph explores a new and fundamentally different approach to evaluating the educational investments that are made by private and public foundations. Such investments typically fund multi-year projects and center around the improvement of one or more key dimensions of the system (for example curriculum, professional development, governance, assessment) that supports districts, schools and teachers. This monograph suggests a new approach to understanding the ways in which and the degree to which such investments are, in fact, "making a difference" and are contributing to the improvement of the quality of education in the nation's schools.

This monograph proposes that the evaluation of externally-funded reform projects should focus on the question of distinguishability. Very simply: Are the people and places who are involved in a funded program distinguishable in significant ways from other similar schools and teachers who are not involved? Are these differences obvious to knowledgeable and skilled observers who are "blind" to the existence of any funded program? And, finally, do the distinctions that are noted signify important differences in substance in either capacity and/or performance?

The approach presented here draws on the work of the mathematician Alan Turing, and, more specifically, on the "test" he invented to answer the question of whether "computers are intelligent." The essence of his test is found in the establishment of a challenge to a skilled but "blind" observer: *Can the observer distinguish between the responses of a computer and those of a human being?*

To understand the ways in which this "Turing Test" approach is different from traditional evaluation approaches, it is first necessary to understand that to date

evaluation has sought to define "the goals" of a program and then translate those goals into measurable outcomes. Instruments are then developed and it is hoped that the measurements made by these instruments will show differences between a program site and non-program site, or, perhaps, between a site when measured in a "pre and post" program fashion. Following this standard approach it is hoped that funded programs will be distinguishable from non-funded programs in terms of the differences of absolute measurements of intended outcomes.

In theory this sounds fine. In practice, this approach rarely leads to clarity or conclusive results. It is simply too hard to define and measure key outcomes, and the error in such measurements is so large that differences in pre and post scores, and / or cross program scores are so large that comparisons are often neither valid nor reliable. In addition, there is often so much "noise" in the experiment that it is very difficult to determine causal relationships between program "inputs" and measured "outcomes."

In this monograph we suggest an alternative approach — one that does not attempt to measure differences through the comparison of absolute measures. Rather the approach suggested here is one that seeks to use expert judgments to make comparisons directly without ever having to define or measure "outcomes" in an absolute sense. At the very least we suggest that this approach is a good complement to current strategies for measuring whether or not the investments made in improving the educational system are, in fact, making a significant difference.

This Monograph

This monograph is a product that is part of our work with the Exploratorium's Institute for Inquiry. The Institute for Inquiry (IFI) is an NSF-funded Center that seeks to help other elementary science reform projects improve their own professional development activities. In designing our evaluation of this effort, it became clear to us that we needed a rigorous but appropriate methodology for evaluating the degree to which and the ways in which the Exploratorium's IFI was contributing to the capacity of these many other existing projects. To their credit the Exploratorium was willing to sponsor our efforts to use this new approach to evaluating the impact of their work.

This monograph is written in three parts. In Part One we introduce and explain the notion of the Turing Test. In Part Two we describe more generally how the Turing Test might be used to create evaluations of educational programs. And in Part Three we describe in detail how the Turing Test approach was used to design a rigorous but appropriate evaluation of the Exploratorium's Institute for Inquiry.

Part One: Understanding the Challenge

There are many public and private foundations who are seeking to make wise investments in the improvement of K-12 education in the United States. Not unreasonably they are asking that these investments be accompanied by an evaluation effort that will help them assess the nature and scale of the benefits that result from the work they are funding.

There is also no doubt that most grants made in education have the ultimate goal of improving the learning of students. But very often the investments that are made in educational reform are focused on elements of the educational system that are far removed from, and indeed "upstream of," the achievement of students.

For example, many educational reform efforts focus on the development of new curricula, the professional development of teachers, or the establishment of high standards and accompanying assessments. These investments recognize that student achievement depends upon providing all students with high-quality learning opportunities, and that such learning opportunities, in turn, depend upon a sound educational system. Hence, funders increasingly believe that their investments must pursue a "change strategy" that is congruent with a "systemic perspective."

Systemic reform is a funding strategy — and a theory of change — that revolves around the belief that the quality of classroom instruction, and hence the opportunity that students have to learn, depends upon a wide number of interrelated system components. And each of these system components (a qualified teacher, a well-designed curriculum, instructional materials and resources, etc.) — is a necessary, <u>but not sufficient</u>, ingredient that is critical in supporting good instruction. When many such critical system components are absent, we would describe that school or classroom as being at high risk of failing. Similarly, we belief that when all system components

are present, and when they come together in a coherent way, the probability of good instruction, and of high student achievement, is greatly enhanced.

Many externally funded reform projects tend to focus on a single system component. That is, a project may undertake a professional development program; another, with different expertise, might work on the development of a new curricula; still another might work with local administrators and teachers to "restructure" the organization of a school. In this way, NSF and other funders support high-quality projects with specialized expertise. In their own way each project makes important but different contributions to improving the overall capacity of the educational system that supports classroom instruction. But, it is also important to note that at the same time, projects such as these, because they are focused on a single dimension of the system, are unlikely to cause large increases in student achievement. More accurately, it could be said that investments in projects such as these seek to build the capacity of the people, and the components of the educational system, which in turn comprise the critical local supports that underlie all good instruction.

The challenge that faces foundations, grantees and evaluators alike is how to evaluate the relative effectiveness of such capacity-building investments. While the ultimate, but distal, goal of educational reform is always increased student achievement, there are many reasons why measures of student achievement are a very poor indicator of the effectiveness of educational grant-making that is, in essence, focused building the capacity of the system.

Because this point is so critical, and so often misunderstood, it is important to spend some time detailing at least some of the difficulties in trying to find direct causal connections between grants in educational reform and increases in student achievement:

- Measuring what students know, especially in terms of higher order skills, is very difficult to do. Most current tests measure basic skills and knowledge. Thus, the current practice of assessment measures only a small part of the spectrum of student knowledge that is deemed as important.

- It is difficult to measure what a student knows and is able to do. But it is even more difficult to measure changes in student knowledge and skills that occur over time. If there is significant error in the test measurement of what a student knows, then there is a much larger error that results when one tries to determine the increase in knowledge (essentially, by subtracting one test score from another).

- Districts, schools, teachers and students are constantly changing, and turnover is often very high. Also, what students know and are able to do is cumulative. What an eighth grader can do is a result of all previous learning, including substantial out-of-school learning. Thus, it becomes a very vague notion at best when one talks about "high achieving schools" or "a classroom being responsible for increased test scores."

- Moreover, there are many different events and programs that influence the goals and nature of instruction; hence, it is very difficult to create a situation where the "impacts" of a single program can be studied in a rigorous experimental fashion. Students and teachers and schools are never assigned to an experiment in randomized fashion. Consequently, it is almost always impossible to find a situation in which changes in instruction, not to mention student knowledge or skills, are attributable to a single intervention.

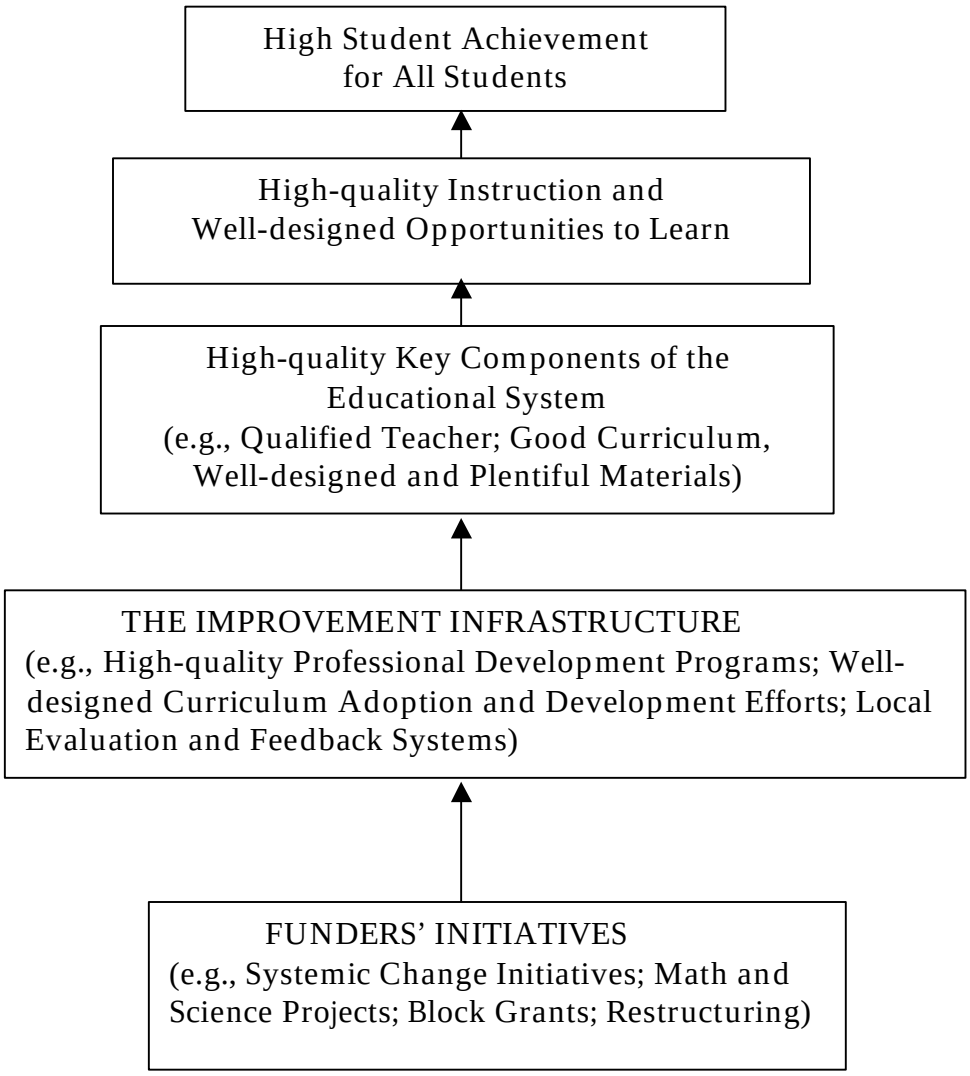
- Educational programs that are funded by external agents often target only one component of the system, and as we pointed out before, this component may well be a necessary but not sufficient condition for educational improvement.[1]

In addition to the above issues, it is very often the case that investments made in educational reform are two or three layers removed from teachers, classrooms and students. Many funded projects that focus on professional development, or new curriculum, or school reform are actually aimed at building the capacity of the system so that the system, in turn, can provide the key supports that are a prerequisite to good instruction. Many funded projects, then, are effectively investments which are aimed at strengthening what might best thought of as the "improvement infrastructure for education."[2]

The logical chain, then, that underlies many investments in education today is illustrated in the diagram below:

[1] It is perfectly possible, for example, for a professional development program to be very successful in increasing the knowledge and skills of teachers and yet never see any changes in the achievement of the students they teach. The reason for this is that other conditions may well mitigate against such improvements. For example, the school environment may be chaotic and unsafe; the instructional materials may be absent; or there may be high stakes tests that do not allow for any change in instruction. So, again, it is very important to understand that high-quality teachers are one of the necessary but not sufficient conditions to create high-quality instruction, and that any given professional development program is only a minor factor in influencing the overall quality of instruction in a classroom setting.

[2] The idea of an "improvement infrastructure" comes from management studies. See, for example, *The Unfinished Revolution* (www.bootstrap.org/colloquium/session_6/col_session_6.html).



In simple language the diagram tells us this: Student achievement depends, in part, on what students learn in classrooms. And what they learn in classrooms depends, in part, on the quality of instruction they encounter there. And the quality of that instruction is itself highly dependent upon multiple critical system components such as the quality of the teacher, the soundness of the curriculum, etc. And the strength of these system components depend, in part, upon the degree to which there exists an "improvement infrastructure" that is capable of providing a continuing process that will upgrade the quality and effectiveness of the key system components that are needed for good instruction.

The Challenge for Evaluation

All of this leads to a deep paradox in terms of evaluation. There is no doubt that those who fund educational initiatives have as their goal the improvement of student achievement. And there is also no doubt that those who work in the programs that are funded must always attend to the distal and ultimate goal of helping students learn more. But it is a profound fallacy to think that it is appropriate or even possible to evaluate the benefits of educational investments by assessing changes in student achievement! As we have argued, most educational grants are long-term investments in improvement infrastructure; they are not direct expenditures for increased student test scores.

There is, then, a very important paradox to be understood here: The goal of investments in educational capacity building is always to increase student learning, but the effectiveness of the investment can not be assessed by measures of student achievement. The failure to understand this reality is not only a technical flaw in the system but actually creates a major obstacle to all existing efforts to improve the functioning of the educational system. Like Nasrudin who looks for the key under the streetlight (not because that is where he lost the key but that is where he can see), we continue to look for evidence of program value in measures of student achievement. This persistence leads not only to a flawed understanding of the value of the investments we are making in educational improvement, but, worse, may ultimately be very counter-productive in terms of misleading people about the value and efficacy of the efforts we are making to improve education.

Football Programs

It is interesting to think in an analogous way about a successful football team. There is no doubt that the success of a football team depends, ultimately, on its win-loss record. But people know that a good football team depends upon the existence of a good football "program." The program is, in essence, the infrastructure that supports the team. A good football program (which involves high-quality coaching, skilled recruiters, a good "front office," supportive fans, and generous boosters) takes years to build. An effort, say, to increase the alumni support for a football program would not be measured by the next year's win-loss record. Rather it would be measured in its own intrinsic terms — i.e., alumni ticket sales — and it would be understood that such support contributes to an overall stronger program.

One might think we are making excuses — that we are saying that it is then impossible to evaluate educational projects in a rigorous way. But this is not at all what we are arguing. Rather we are simply asserting that evaluators must then find a way to rigorously evaluate the benefits of investments in ways that are appropriate to the nature and likely outcome of those investments. Currently, we would argue, that evaluation, because of short-term political pressures, too often evolves into a pseudo-rigorous approach that relies on inappropriate outcome measures. By contrast, we are seeking a rigorous way to evaluate appropriate outcomes. We are looking for a scientific way to assess the degree to which and the ways in which educational investments are, in fact, leading to increased system capacities for continued program improvement.

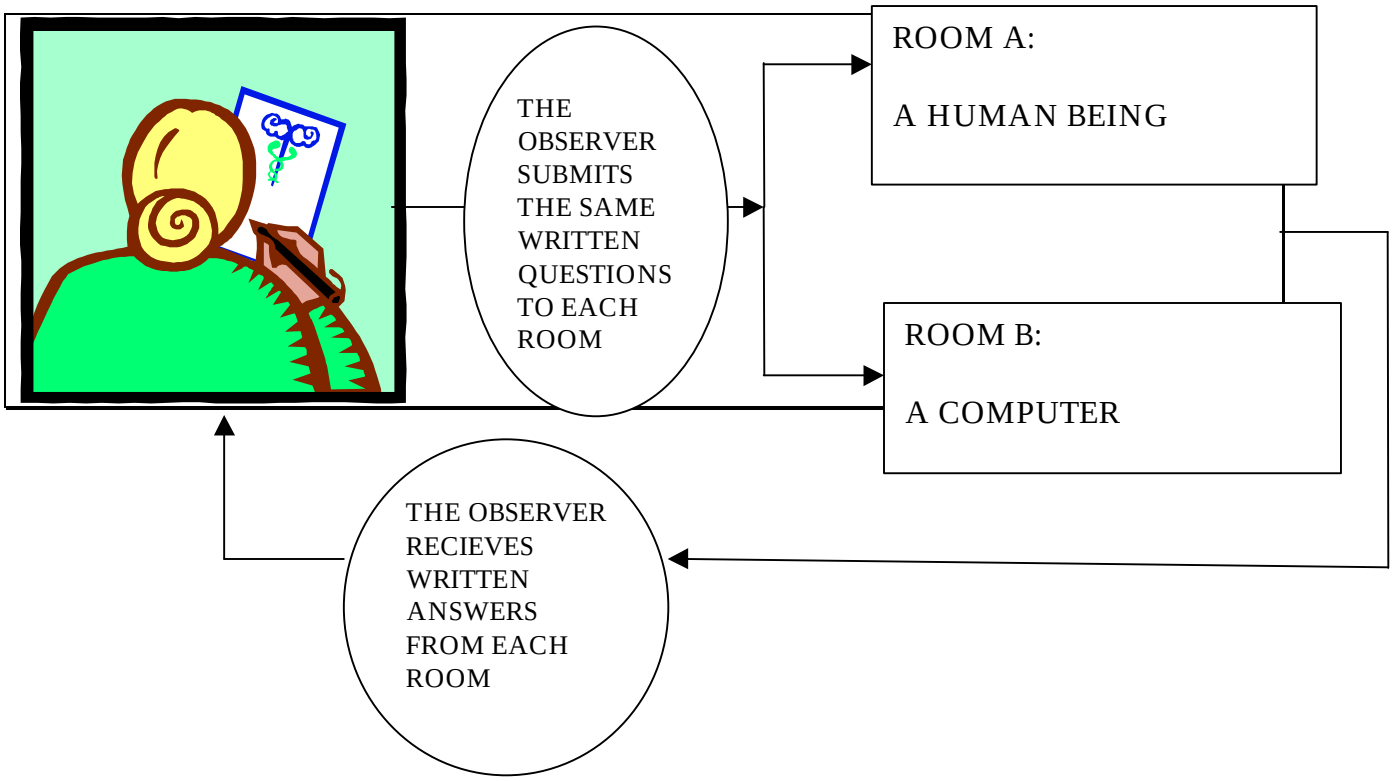
Part Two: The Turing Test

In 1950 the mathematician Alan Turing published a paper entitled "Computing Machinery and Intelligence."[3] In this paper he suggested that machines could achieve "intelligence," and he came up with a simple test of this proposition. He suggested that there might be a simple operational test of whether or not computers were "intelligent." He suggested a kind of *gedanken* or "thought experiment" in which one might imagine a thoughtful observer sitting outside of two different rooms, each with a closed door. In one room there was a human being, in the other room a computer. On the door of each room there was a slot through which one could pass a written message.

The challenge to the observer was simple: Try to figure out which room held the computer, and which the human being. The rules were simple. The person outside obviously could not open the doors and look inside. Rather they were allowed to formulate questions, and then submit them in writing through the slot in the door. And written answers would emerge from the same slot. The person outside the room could then try to assess the answers that came from each room and determine in which room the computer was located.

[3] See Paul Strathern's *The Big Idea: Turing and the Computer*, page 86 (New York: Anchor Books, Doubleday, 1997).

AN EXTERNAL INTERVIEWER



Turing called this an "imitation game" because the computer was seeking to imitate the human response. This "game" later became known as the Turing Test. If the person outside the room could NOT correctly guess which room held the computer, then the computer would be doing well. So, Turing asserted, if the outside observer could not consistently and reliably distinguish between the written responses of the computer and the human, then, Turing asserted, the computer had achieved intelligence.

It is important to note right away that the key to this test is the establishment of a comparative situation in which an intelligent observer is asked to use all of his or her knowledge and skills to distinguish one alternative from the other. Or looking at it from the other side, those who design the computer are trying to achieve "in-distinguishability" in the eyes of an intelligent, skilled but "blind" observer.

It is also very important to understand that the power of the test depends upon the level of knowledge and skills of the person outside the room. The test allows this person to ask whatever questions they like. Thus, their challenge is to find good questions — ones that will be deliberately aimed at exploring features that are likely to yield differences when answered by a human or a computer. (By examining the questions that are asked, one can also use the Turing Test to reveal the ways in which the interviewer thinks that computers and humans are most likely to be different.)

The key here for our purposes is to see that the Turing Test creates a situation that explores the degree to which and the ways in which two situations are indistinguishable (or, conversely, distinguishable). If an ordinary person, using all the cleverness they can, is not able to distinguish — in a blind test — between the responses of a computer and a human being, then we have to conclude that the computer and the human are equivalent — at least along the dimensions that are explored.

The Turing Test Approach in Education

We believe that the logic of the Turing Test actually underlies much of what evaluation seeks to do. More than anything else, the funders who provide resources in the form of educational grants want their investments to "make a difference." They want the work that is supported by their grants to make the world a better place — in a distinguishable way. That is, at some level, funders want the initiatives they fund to make a noticeable, tangible and significant difference in the educational system, and ultimately in the lives of people. Note that there are two important criteria here. One is that the differences generated by the work of the initiative be large enough to be distinguished, and the second is that the distinguishable differences be found in dimensions that are judged to be important.

The Turing Test approach hinges on the idea of direct human comparison, and not on intermediary measurements. The Turing Test says that if a skilled observer can not

reliably distinguish between two anonymous alternatives using normal methods of interrogation, then there is no significant difference between those two alternatives. Hence, we would argue along with others that an intelligent and skilled search for distinguishability is a key underlying principle of evaluation, and, indeed, of all science.

Double Blind Studies

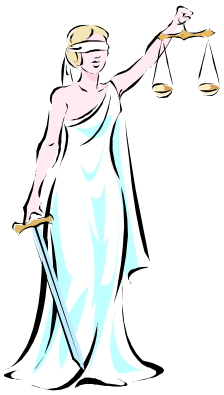
" Double blind" studies are often used in medicine to determine the efficacy of a new drug. In this case two different groups of patients receive two different treatments — one gets the experimental drug; another gets the placebo. Importantly, neither the doctor nor the patient has any knowledge about which drug is going to which patient. Then the question becomes whether or not the two different groups are statistically distinguishable from each other or not. (In such studies random selection is used to maximize the probability that the differences that are observed are most probably due to the treatment itself and not some other factor.) But the point is that 'double blind' studies are but one specific example of the more general Turing Test approach.

Recalling the initial challenge we posed at the beginning of this article, we now pose the following question:

> ***How can we use the Turing Test, and the basic idea of asking a knowledgeable but "blind" observer to distinguish between two alternatives, to assist us in evaluating investments that seek to build educational capacity and enhance the local improvement infrastructure?***

As we said before, funders basically want to know that their grants, which they see as investments, are "making a difference." Too often, as we have stated, the desire to check that their grants are "paying off" leads to evaluations that seek to measure the impact of the grant on distal and often inappropriate outcomes (such as student achievement). What is being suggested here is that evaluation can be designed to create a more immediate test of whether or not the grant has made a difference — by seeing if the results are clearly visible to a "blind" but skilled observer.

.....



It is interesting to note that "justice" is often portrayed as a woman, blindfolded, with a scale in her hands weighing two alternatives impartially.

.....

The Turing Test approach suggests a fundamental departure from current evaluation approaches. The Turing Test suggests that evaluation might be better off by establishing distinguishability as a first-order outcome. Alan Turing saw how difficult it would be to measure, in an absolute sense, the intelligence of a computer, or of a human for that matter. Intelligence is multi-dimensional, dynamic, and not easily assessed. Consequently, he rejected the standard evaluative approach: That is, he did not seek to find a way to measure the intelligence of a human or a machine in an absolute sense. He did not belief that he could develop an intelligence "thermometer" which would accurately measure intelligence — and then use those measurements to determine the level of intelligence of the computer and of the human. He did not use "pre and post" measures, nor did he even compare measures of intelligence given to one room or the other. Rather, he decided to examine the degree to which computers were distinguishable (or not) in the eyes of another intelligent human being. The key idea here is that the measurement instrument that is relied upon is expert human judgment — a remarkably powerful and sophisticated instrument! The question of measuring intelligence is, in the absolute, a very complex undertaking, but the Turing Test made it a relatively straightforward comparative and concrete procedure that uses

expert judgment to compare two different intelligences. The experiment also has the advantage of being easily understood by both technical and non-technical audiences. The logic is compelling, and the test is blind and therefore clearly rigorous.

We propose, then, that the Turing Test suggests a new way to evaluate investments that work in similarly complex situations. We suggest that expert observers can be used to examine the distinguishability of two or more complex situations, some of which are influenced by the funded program while the others are not so influenced.

In complex situations it is far easier to evaluate effectiveness by having a situation that is inherently comparative rather than relying on the comparison of absolute measures.

For example, in developing the boats for the America's Cup yacht races, most syndicates seek to have "two boat" campaigns. That way they can make changes in the design of one boat and compare its relative speed to the other companion boat — the design of which they have not changed. There are many factors that affect the absolute speed of a boat (like wind speed, wind angle, wave height, and wave frequency). And, to complicate things more, these factors vary continuously, so that it is nearly impossible to assess small changes in boat design without a matched comparison that is sailing in the same dynamic conditions.

We believe it is possible to use similar simple tests in assessing the value of investments in educational reform. Here in simple steps is how we imagine Turing Tests done in educational settings.

1) **The investment or project to be evaluated must make assertions about the degree to which and the ways in which they believe institutions, programs, practices, and/or people become distinguishable as a result of the work they do.**

2) **The evaluator then formulates a Turing Test — a procedure in which skilled researchers are asked to examine two or more situations and make comparative judgments about their salient differences. Both researchers, and those working in the comparative situations, are "blind" to the test.**

3) **Reviewers then analyze the data gathered by the researchers, as well as the ratings of the researchers, to independently see if the situations they studied are, in fact, distinguishable, particularly along critical dimensions.**

In what follows we flesh out these abstract steps by describing an experiment we did recently in evaluating the impact of the Exploratorium's Institute for Inquiry — a program designed to build the capacity of elementary science education reform efforts.

Part Three:
A Case of the Turing Test in Educational Settings —
Evaluating the Impact of the Exploratorium's Institute for Inquiry

Background

The Institute for Inquiry (IFI) is a program funded by the National Science Foundation (NSF) and conducted by the staff of the Exploratorium, an internationally known science museum located in San Francisco, California. The purpose of the Institute is to assist other NSF-funded elementary science reform efforts, particularly Local Systemic Change projects (LSCs). More specifically, IFI seeks to help the LSCs improve their professional development programs so that they, in turn, help teachers use an inquiry-based approach in their own classrooms.

The Local Systemic Change projects are located around the United States and have the mission of improving elementary science education by providing all the teachers within a targeted district with extensive professional development. There are over two dozen funded LSC programs specifically funded to promote reform in elementary science education. These LSCs seek to help local districts implement high-quality elementary science programs that include well-designed curriculum and a teaching force skilled in inquiry-based instruction. The Exploratorium IFI program seeks to help Local Systemic Change initiatives improve the quality of the professional development they offer their local teachers — especially in the area of inquiry-based teaching and learning.

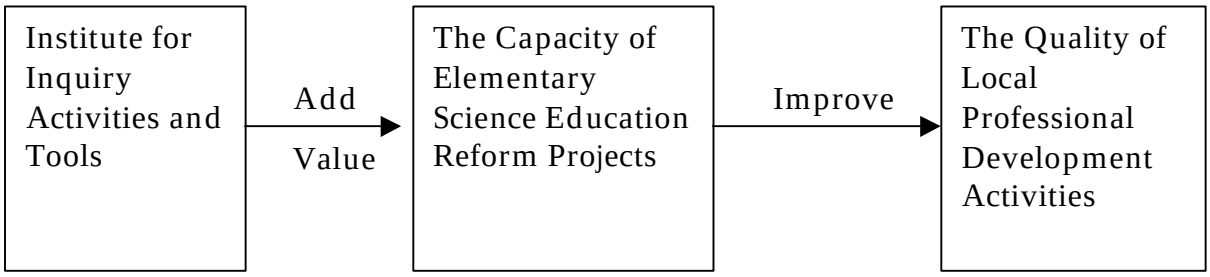
If successful, the IFI program would lead to the following kinds of outcomes:

- LSC leaders and professional developers who have a much deeper personal understanding of inquiry teaching and learning,

- Improved LSC offerings that reflect a greater coherency in design and more sophisticated approach to inquiry teaching,

- Greater connections between the LSC and other LSCs, as well as the Exploratorium, around a shared interest in inquiry teaching and learning

The logic of the Exploratorium's Institute for Inquiry is outlined below:



Designing a Study Based on the Turing Test

We believe that there are several reasons that this program offers a good candidate for a Turing Test approach.

- One is that IFI is seeking to build leadership capacity and influence the design and practice of other reform initiatives. In this sense the outcomes that can be expected from this program are quite distal from classrooms and students.

- The appropriate assessment of the IFI program involves assessing the degree to which the program is "adding value" to existing reform efforts, and then through that assistance helping those local reform efforts do a better job of their local professional development activities. These are benefits that are very hard to measure in an absolute sense.

- The whole situation is very complex. As with the America's Cup boats, it is very difficult to tell how much difference — and what kind of differences — IFI is making with the elementary science reform efforts that it serves. Hence, it is important to compare reform efforts against each other, since they all face similar kinds of adverse winds, currents, and variable weather conditions.

The program is clear about the outcomes it wishes to achieve — leadership vis-à-vis inquiry, and improved professional development design and practice — these outcomes are not easy to measure in an absolute sense. But they are possible to compare across different reform projects.

Setting Up the Turing Test

We asked the following question:

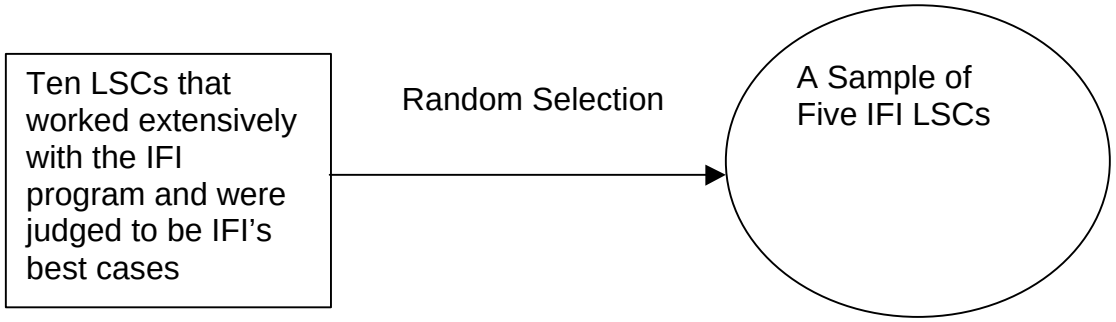
> ***In what ways and to what extent are the LSCs that IFI has worked with distinguishable from the LSCs that IFI has not worked with, especially in those dimensions that are most important to the IFI program?***

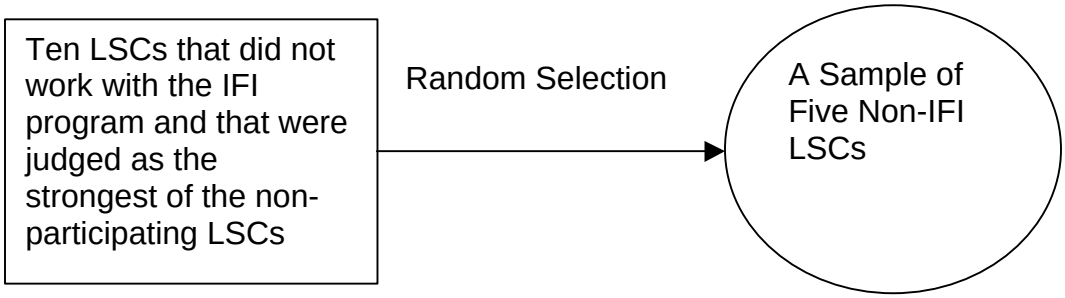
That is, we asked about the degree to which — and the ways in which — IFI clients would be distinguishable from the non-IFI clients. If, in fact, outside expert observers could find no differences between the two, then it would be hard to justify the NSF investment made in the IFI program.

The Procedure

Since it is not possible for outside reviewers to visit large numbers of LSCs all over the country, we selected the following sample of LSCs for our comparative test:

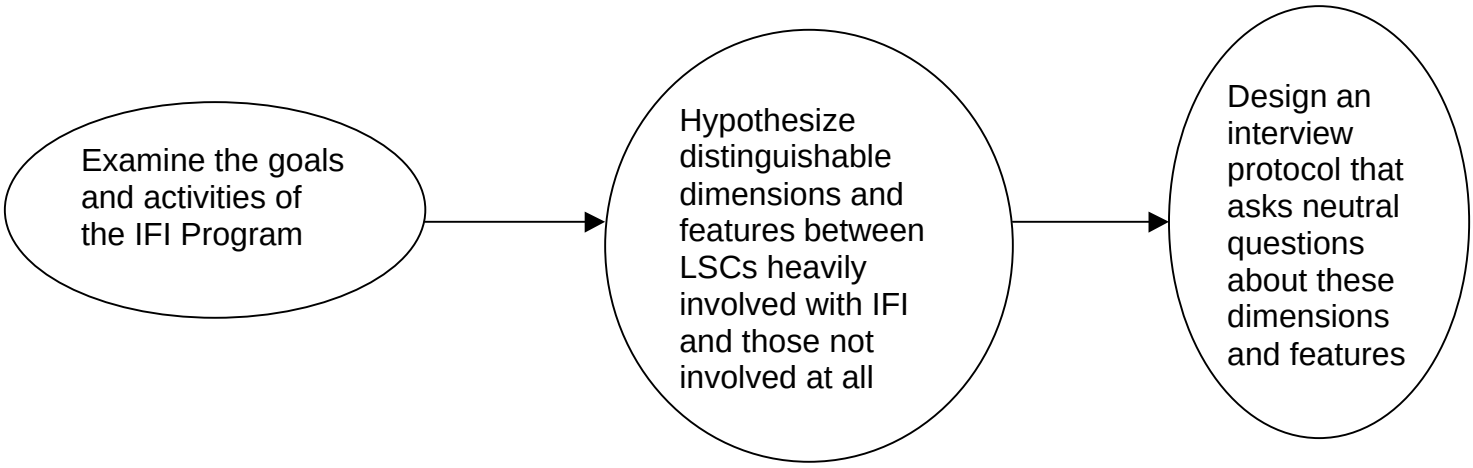
1) We selected two samples of LSCs — one was a group of IFI clients, and another a group of LSCs not involved with IFI. We asked IFI to identify the "top ten" of its clients and from that sample we randomly selected five to be included in the Turing Test. Similarly, we identified what we believed to be the ten strongest and most mature of the non-IFI LSCs, and we randomly selected five from that group.



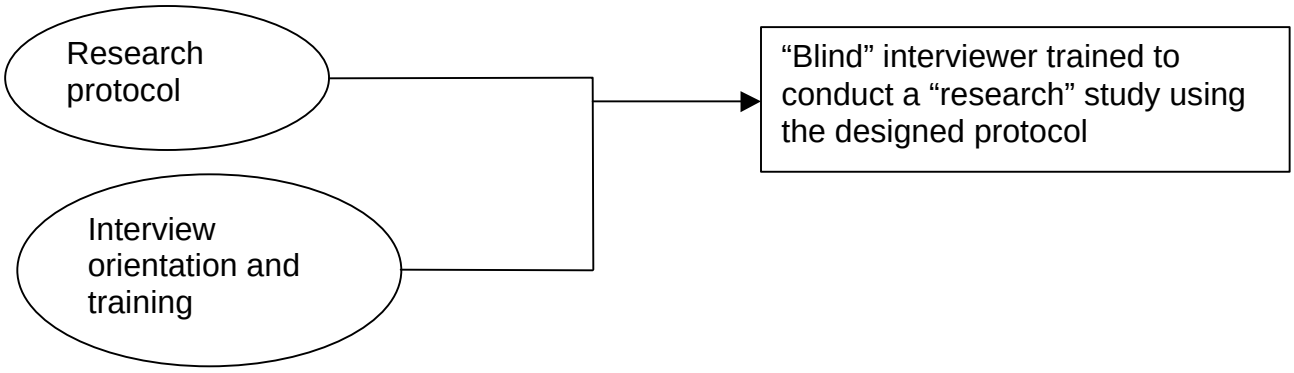


2) We created a protocol that was designed to be used as a template for interviewing the leader of each LSC. This interview protocol[4] asked about certain dimensions of the work of the LSC, including its stance toward inquiry and its approach to designing professional development. As with any Turing Test, the questions were selected to maximize the probability of finding differences between the two samples, but in no way did the interview protocol use language or terms that would be a "tip-off" or "code" to the IFI LSC participants.

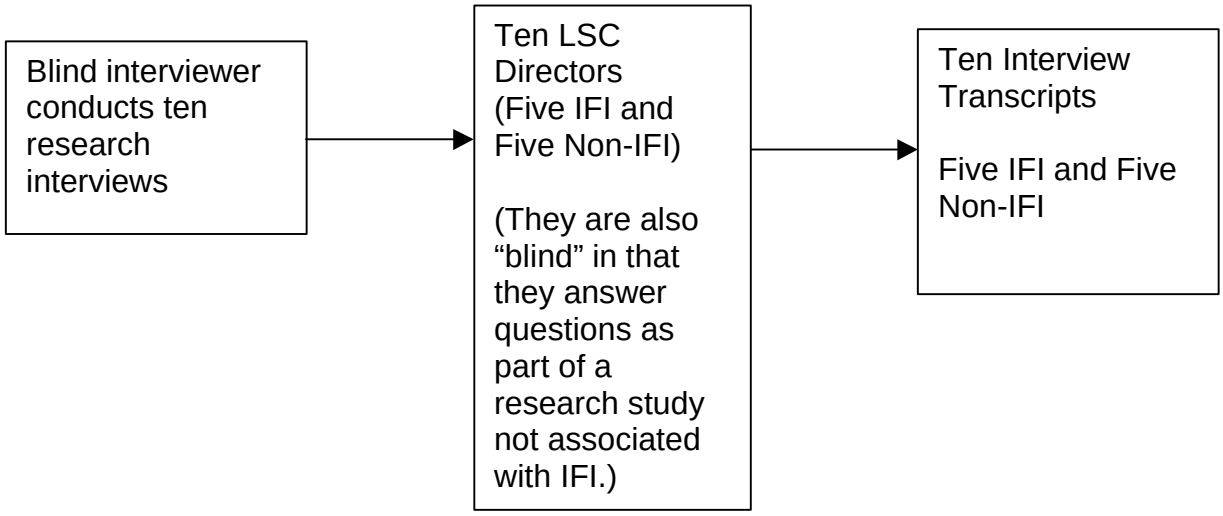
[4] See the Appendix for the interview protocol.



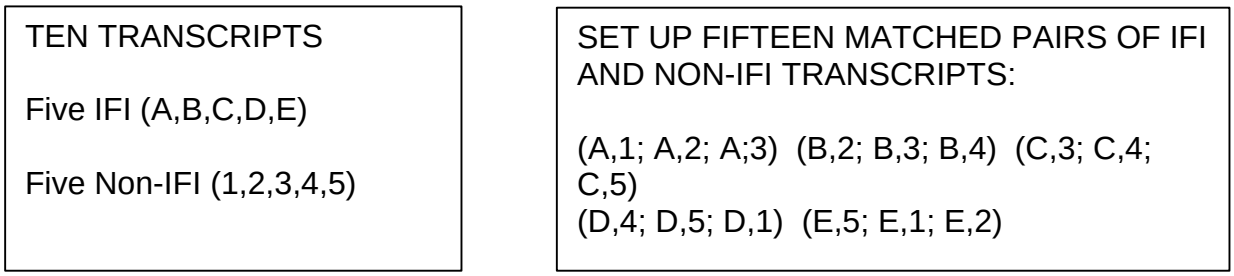
3) We trained an interviewer to ask the questions on the protocol. She was told that the interviews were part of a research project that was investigating the ways in which different LSCs thought about their work and designed their activities. She did not know that the interviews in any way involved the Exploratorium or IFI, and, of course, she did not know that there was any difference between any of the ten LSCs that were part of the study.



4) The interviewer conducted an interview with the leader of each LSC included in the sample. The interviews were tape recorded and transcribed.

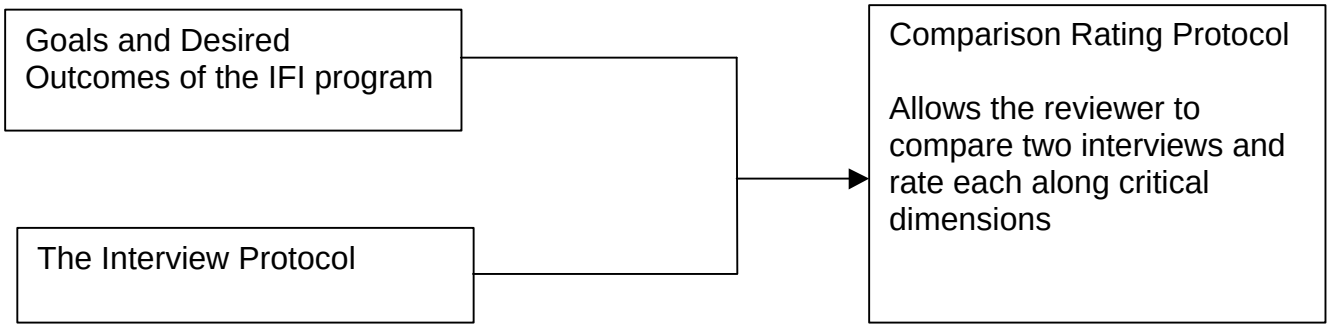


5) The transcribed interviews were then assembled and matched randomly in pairs — with one IFI LSC interview couple with one non-IFI LSC interview in each pair. In all we set up 15 matched pairs.

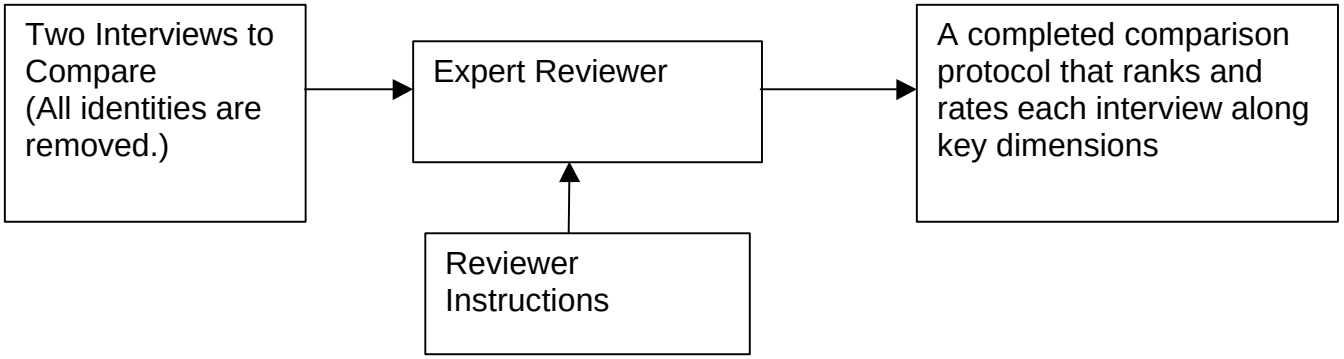


We created a comparison protocol by which an outside reviewer could read two interviews and rank the two along critical dimensions and according to established criteria. The comparison protocol asked reviewers to make inferences and judgments about the project, and the leadership of the project, based on the responses of the interviewees contained in the transcript.[5]

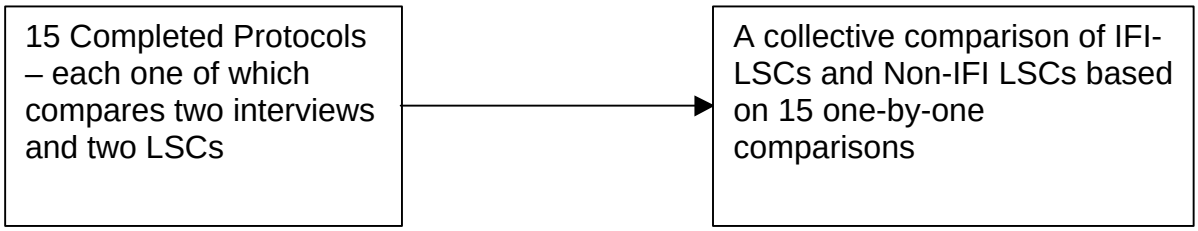
[5] See the Appendix for the comparison protocol.



We then identified and recruited ten expert reviewers. These reviewers were skilled in science education, professional development, and inquiry-based teaching and learning. We sent them each a pair of interviews and asked them to compare the two interviews along certain dimensions. The reviewers were not aware that they were part of any kind of evaluation; rather, they understood that they were involved in a blind study of some sort that involved the LSCs. The comparison asked interviewers not only to judge which interview was superior on a given question, but also to rank each interview on a five-point scale. (Some reviewers ranked two sets of interviews.)



We then collected and compared the completed interview comparison protocols. This provided us with 15 different cases where an IFI-involved LSC was compared against a non-IFI involved LSC.



It is important to note that this test comprises what is, in essence, a triple blind study:

- The interviewer did not know which LSC was which, or even that this was an evaluation effort involving IFI and the Exploratorium. Hence, this eliminates the natural tendency to probe certain answers and seek for more information that would illuminate the work of the Exploratorium or the issue of inquiry more generally.

- The interviewees did not know that this study was connected with or part of any evaluation. Rather they answered the questions on the basis that they were part of a more general research study and as part of that they tried to inform the interviewer about their thinking, their design, and their practices.

- The reviewers similarly did not know that they were part of any specific evaluation study. Rather they simply knew that they were hired to judge in a blind fashion interviews with different LSCs.

An Analogy: A Series of Horse Races to Determine the Best Stable

Supposing you wanted to know which racing stable did a better job of training race horses. This is tricky since race horses vary considerably and conditions also are not steady. One way to make the comparison would be to draw randomly a sample of five horses from each stable — and then race them in pairs. You might think of the comparison of these 15 matched pairs of interviews as 15 horse races. The goal of these races is to see whether one stable is better than another stable at producing fast race horses. The 15 horse races involve a total of five horses from one stable, and five horses from another stable. They then race each other one at a time, with each horse racing three other horses from the other stable. (Due to limited resources in our IFI study we did not have each horse compared against each other horse — which would have been optimal.)

The Results — Looking at Rankings

One way to decide whether the LSCs that IFI has worked with are distinguishable from those that it has not assisted is to compare the winning of "individual horse races." For each question on the comparison protocol we asked reviewers to "choose a winner," deciding between the two different interviews they read for a number of specific questions.

In the chart below we compare the number of times that the IFI-involved LSC was judged to be superior to the non-IFI supported LSC.

THE COMPARISON QUESTION	*IFI-INVOLVED LSCs*	*Non-IFI LSCs*	*Grand Total*
Overall, which project would you say has most benefited from sources of outside support?	13	2	15
Overall, which project do you feel has the deepest understanding of and commitment to inquiry?	10	4	14
Overall, which project do you feel has the deepest understanding of and commitment to leadership development?	10	4	14
Overall, which project do you feel has the deepest understanding of and commitment to professional development design?	10	4	14
Overall, which project do you believe has the strongest vision for science teaching and learning?	10	5	15
Grand total	53	19	72

It is clear that the IFI involved LSCs are distinguishable from the non-IFI LSCs — at least along these questions. They win "the matched pair horse race" at a ratio of more than two-to-one. Collectively, along these key questions, the IFI LSCs "win" 53 out of the 72 comparisons.

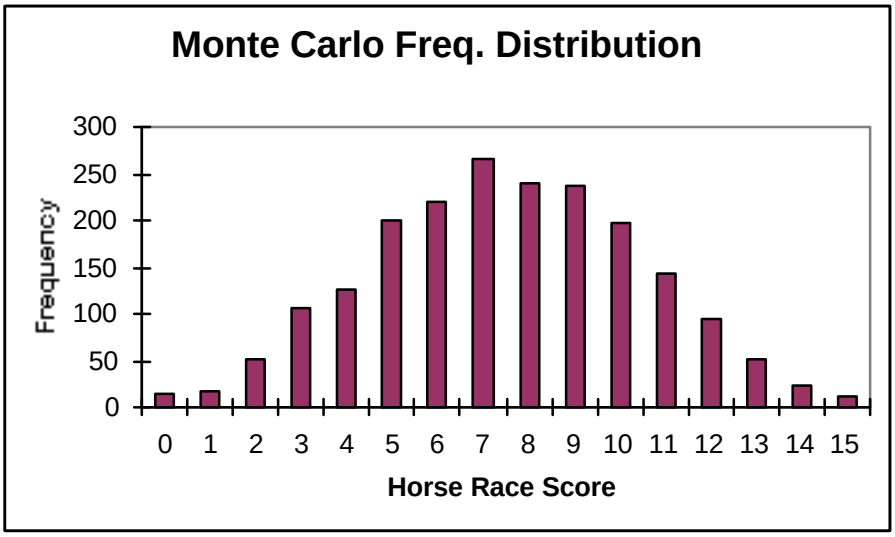
It is interesting to ask whether these results are "significant" statistically. That is, what are the odds of one stable winning 10, 11 or 12 races out of fifteen, if, in fact, the stables were equal?

Statistics

Basically, the rhetoric of the horse-race argument is this: Assume that these two projects categories (IFI and non-IFI) are equally "good horses." Under the conditions of the evaluation design (partial crossing with raters), and the assumption of equal quality between the horses (project categories), what is the probability that we would observe, say, a 9 to 6 result? What about a 10 to 5 result? How extreme does the score need to go before we can claim that the probability of observing these results, given equal horses, drops below a reasonable threshold?

Rather than try to work out the probability model algebraically, an alternative is to use a computer simulation. Basically, one can assign the ten projects a random score. Then, for each of the actual project comparisons in the evaluation, a "rater" can judge which was better (based on the random

numbers). This results in 15 horse-race judgments under the "null" model that all projects were drawn from the same quality group. If we replicate this equal horse race 2000 times, we can then count how often we see a score of 15-0, 14-1, 13-2, and so on. This gives us an approximation (and a very good one) of the probability distribution we're after, and we can examine it to see where the magic 90% or 95% cutoffs are.



Wins	Freq	Percentage	Cumulative Percentage
0	14	0.7%	0.7%
1	16	0.8%	1.5%
2	52	2.6%	4.1%
3	106	5.3%	9.4%
4	127	6.3%	15.7%
5	200	10.0%	25.7%
6	221	11.0%	36.8%
7	267	13.3%	50.1%
8	241	12.0%	62.2%
9	238	11.9%	74.1%
10	198	9.9%	84.0%
11	142	7.1%	91.1%
12	93	4.6%	95.7%
13	51	2.5%	98.3%
14	24	1.2%	99.5%
15	11	0.5%	100.0%

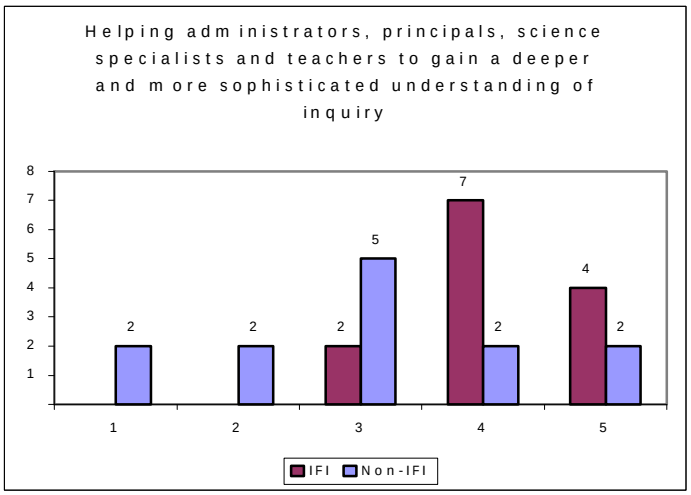
The graph and the chart shown above, then, tell us the following in terms of the probability that the IFI projects are, in fact, superior: Every time IFI wins 10 out of 15 "horse races," we know that happens only nine times out of 100 by chance. And when IFI wins 12 out of 15 comparisons, that happens only five times out of 100 by chance. Thus, while not statistically

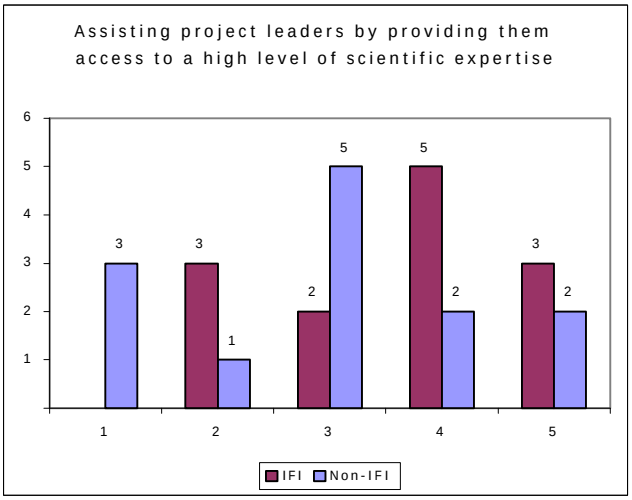
significant at the 95% level, one can say that there is very strong "probable cause" to believe that, by comparing the match-ups, IFI-supported districts, are, in fact, distinguishable from their counterparts.

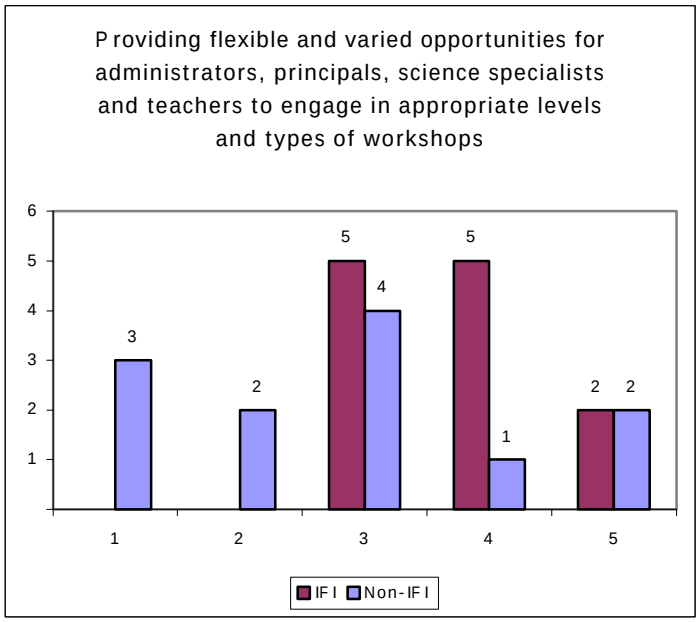
The Results — Comparing Ratings

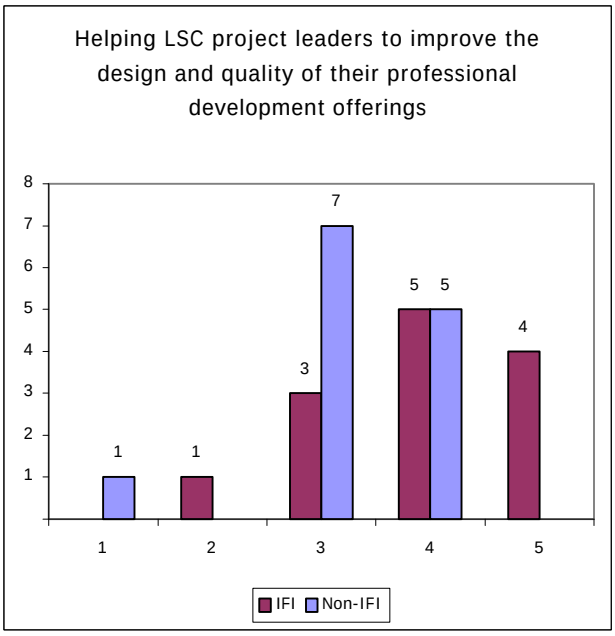
We also asked reviewers to rate each interview (from one to five) along certain key dimensions. The comparison of the reviewer ratings are shown in the graphs that follow.

THE TURING TEST: 15 EXPERT COMPARISONS OF IFI AND NON-IFI DISTRICTS[6]



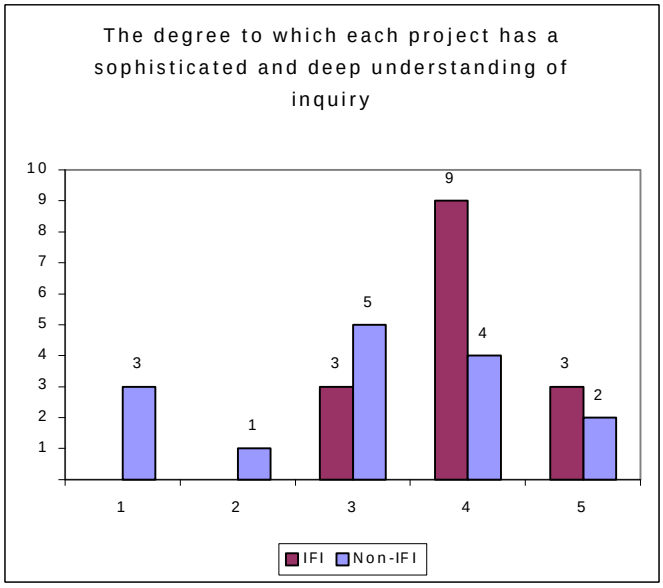


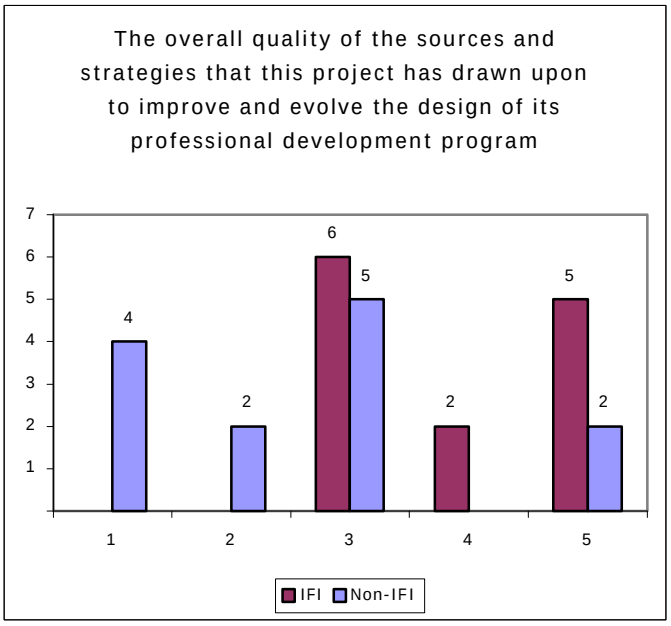


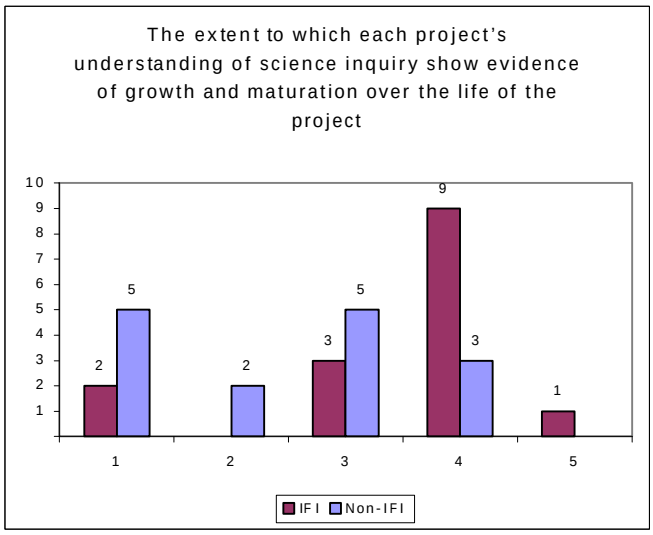


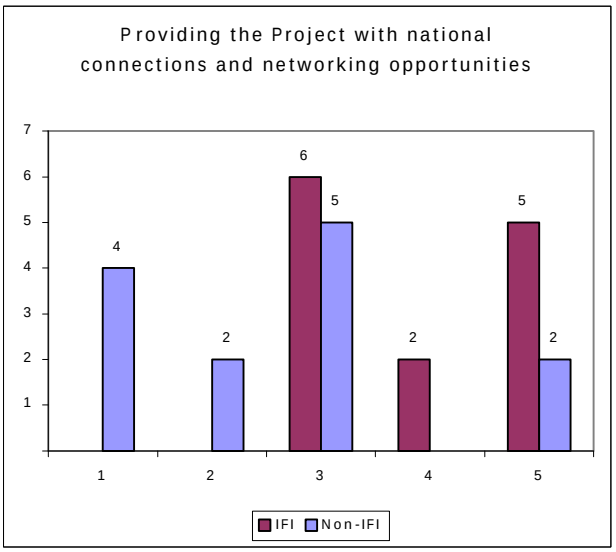
[6] These graphs depict the ratings of 15 expert reviewers as they compared the transcripts of interviews with LSC project directors, half of whom were involved with the Institute for Inquiry and half who were not. The vertical scale represents the number of raters assigning the project a given rating. The horizontal scale refers to the quality of the project's work. The labels on the horizontal scales vary but in all cases "1" represents the low end of the scale, while "5" represents the high end of the scale.

THE TURING TEST: 15 EXPERT COMPARISONS OF IFI AND NON-IFI DISTRICTS









Analysis

The analysis of the "horse race" did not produce inferentially significant results at the 95% level. (This may well be due to a very small sample size.) But when one looks at the differences in the ratings judges gave individual questions, then the average scores of the items in a particular category showed statistically significant differences. This is not entirely surprising because the average scores provide a lot more information than a simple binary "win/lose" item, and are thus more precise.

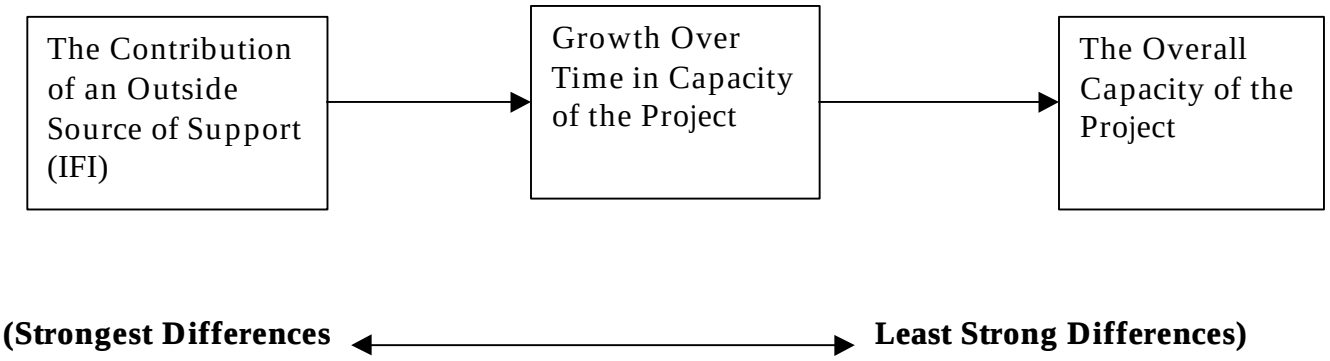
Statistics

Just for completeness the table for the Analysis of Variance is given for five different key questions below.

Category	*p-value (2-tailed)*	*p-value (1-tailed)*
Strength of Vision	***.229***	***.115***
Role of Inquiry	***.004***	***.002***
Strength of Leadership	***.053***	***.026***
Sophistication of Professional Development	***.164***	***.082***
Degree of External Contribution	***.130***	***.065***

In this case, we can make a fairly strong claim that the IFI projects rated significantly higher than the non-IFI projects in four out of five categories (all but Strength of Vision).

It is interesting to note that the differences between IFI and non-IFI supported projects were strongest in the questions around the contribution of outside sources. They were next strongest in the change over time, and least strongest in the absolute value of the differences. This is completely in accord with the way in which IFI is contributing to these projects.


The Contribution of an Outside Source of Support (IFI)
Growth Over Time in Capacity of the Project
The Overall Capacity of the Project
(Strongest Differences
Least Strong Differences)

SUMMARY

This example, we believe, provides an existence proof for a different approach to evaluating investments in educational improvement. It points out how it is possible to ask and answer the question of whether such investments "makes a difference" — one that is distinguishable by experts and that is also educationally significant. And the Turing Test approach we describe here not only answers that question rigorously, it also answers the question appropriately.

It is clear and even inevitable that in this era of accountability that all funded projects will themselves be ultimately held accountable. But we think it is very important that accountability itself be held accountable. That is, evaluation of investments in educational improvement should meet certain basic criteria in order to have legitimacy. The evaluation of funded projects should focus on appropriate outcomes; they should be independent and rigorous; the inferences that are drawn from their results should be scientifically valid; and they should meet a criteria of simplicity and common sense. We believe the Turing Test goes a long way to meeting these criteria.

The example we have provided here shows how evaluation can be used to rigorously assess the contributions of a very good program. Had the Exploratorium's Institute for Inquiry been judged on the basis of student achievement, or had we tried to measure in an absolute sense the capacity of districts to provide teachers with high-quality inquiry-based professional development, we might well have failed to capture any of the program's contributions. Rather, we have used the Turing Test to focus on issues of distinguishability, and answered the common-sense question of whether the program is making a significant contribution to the improvement of science education.

We have used the Turing Test in other situations[7] and we believe there is a rich opportunity to expand the conceptual and procedural underpinnings of this approach. We invite others to join us in that task.

[7] Please see, for example, our December 2000 report: *Inverness Research Evaluation of the MARS Project: District Comparison – Turing Test Report*, on the Mathematics Assessment Resource Service website (http://www.nottingham.ac.uk/education/MARS/eval/dblind.htm).