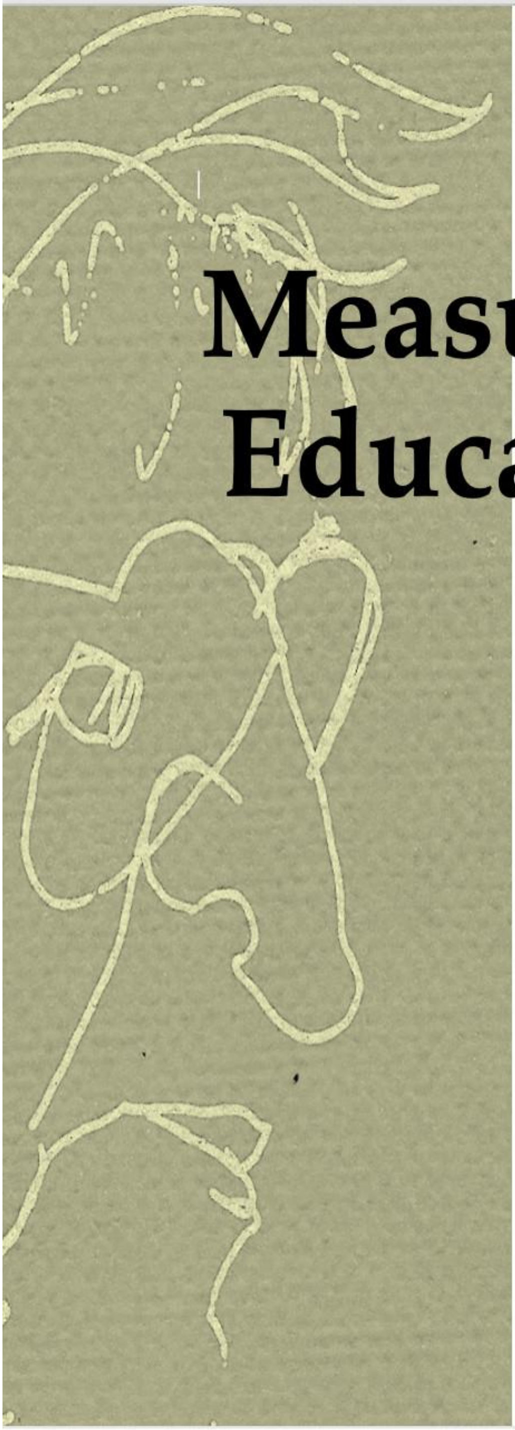


Measuring Education

memoir

Bob Stake

Contents

Measuring Education 1
1965 9
Mapping Knowledge 10
1967 11
Hybrid Seed Corn 11
1968 15
A is A 15
1969 17
Information and the Ever-Normal Granary 17
1971 26
Helping The Disadvantaged 26
1972 29
Beef Stew 29
1973 40
Genesis 40
An Anthropometric Station 47
A Hawk and A Handsaw 60
Voices 67
Responsive Evaluation 82
1974 87
Navigation toward the Bells of Villingen 88
1975 102
SAFARI and Legacies 102
Resigning from the 111
ESAA-SDC Meta-Evaluation Panel 111
Teachers versus State Testing 115
1976 135
Yearning Power 135
On Seeing and Measuring 138
Tom Hastings, Mentor 141
1977 143
External Review of the TORQUE Materials 143
1978 148
A Case-Studies-in-Science-Education Footnote 148
Use of Research Time in One Multiple-Case-Study 151

Images of Governing 153
1979 156
Education Unseen by Politicians 156
1982 161
On Being and Becoming 161
Mail Survey Franking 164
1983 167
Anarchists Against Evaluation 167
1984 169
Art Education and Critical Thinking 169
A Brazilian Field Trip 173
1986 180
Failure to Evaluate 180
1987 187
American Education, 1987 187
The Ionesco Clock 190
Programs that Work 194
1988 196
Midlife Learning about Program Evaluation 197
1989 202
Winning 202
1990 205
Catching the Wry 205
1992 210
Presumptuous Indicators 210
Shadow Study of a Sixth Grader 213
1994 218
Jack Easley, Insight Finder 218
1997 222
Evaluation of "Reader Focused Writing" 222
for the U. S. Veterans Benefits Administration 222
1999 230
The Misanthropy of Effect 230
2000 232
Assessment in U. S. Education 232
Dreaming His Dream? 243
Evaluation Research Strategies 247
2001 251
"That'll Learn 'em!" 251
2002 255
Conception 255
2003 257
How Far Dare a Program Evaluator Go 257

Toward Saving the World? 257
2006 267
Criteria for Naturalistic Case Studies 267
2009 269
Criterial versus Experiential Evaluation 269
Criticism and the University 278
2011 297
Graduation Rates 297
2013 298
Freedom from the Rubric 299
2018 301
Right and Wrong 301
Terry Denny, Story Teller 304
Those Not Chosen 307
2019 316
The Diptych of Research 316
Coping with Voids 320
The Meaning of Meaning 323
2020 327
Making Research More Humanistic 327
2021 329
At 90, In the Rear-View Mirror 329
2022 332
A Reconciling 332
Some Kinds of One 340

Measuring Education

In a way, my life's work can be summarized in short writings. At least, from the time I arrived at the University of Illinois until the present, such a summary is this book. It is a selected collection of essays, the criterion being what best entertains and what tells of my drift from a criterion-referenced measurements man to an experience-watching story-teller. As with many jogs along the way, it has been not so much a change because new talents and insights emerged but more because old ones wore thin. As with the settlers of the West, it was not so much an opening of new territories as running away from the grasps of old.

In 1949, returning to Howard Hall after first meeting me, Bernadine told her room-mate Lela Mae, "He's such a show-off!" Later, Bernadine married me. If I did choose well this collection of short papers, it's because I found words that showed me off.

I feel that that the arguments I made fifty years ago about measurements in education are different from what I have been saying recently. Still, some readers will say that Bob kept saying some of it over and over. Yes, I lost some rationality and gained some humanity. I lost a lot of optimism and gained a little self-confidence; but a thirty-year-old's consciousness is still driving the ninety-five-year-old fingers over the keys.

I started thinking of this as a collection in 1974 when on a family sabbatical year in Sweden and England. I was reflecting on ten years as the Assistant Director of the Illinois State Testing Program, an aide to Director Tom Hastings. He recruited me on the docks at San Francisco, thinking I might fulfill his desire to link achievement testing better to classroom instruction. But the big fascination when I arrived in Urbana was Curriculum Evaluation, and Lee Cronbach and Jack Easley and others pulled me toward what we first thought was a new use for testing. If I drifted, that's when the drift began.

At first, I called this collection, "Through a Measurement Darkly." Then, finding that I had 26 papers first-drafted or contemplated, I called it, "Half a Deck," alluding to the fact that some of us appear to operate with less than a full deck. I never finished all of that particular 26, but I kept writing rants and repartee, which later some people called "blogs," and now here I have picked out more than 52 that I like, a full deck, probably not.

The titles were obscure allusions, such as "Genesis," "Beef Stew," "A Hawk and a Handsaw," and "A Reconciling." The unspoken subtitles were more informative. Here are the first 13 of them: Learning Contexts, Skill Hierarchies, Concepts of Education, Roles of Measurement, Observation, Individual Differences, Reliability, Validity, Metaphor, Generalizability, Formative Evaluation, Summative Evaluation and Responsive Evaluation. In test talk, pretty staid. I thought I might title the collection: "Measuring Education."

With devious mien, I saw it as an attractive title, because I was learning again and again that education, whether a human property or a social agency, could not be measured. Pressure points could be tapped and stories told, but no clustering of measurements could satisfactorily tell the nature or generativity or unwell-being of education. I try to make this point many times in the pages of what ultimately was becoming a memoir.

With a few essays completed in a summer in Boulder, and all of 26 planned, Deborah Laughton, my SAGE and later Guilford editor, and my students-turned-friends, Gene Glass, Stephen Kemmis and Saville Kushner, urged me to make it a book. But the missing chinks in this planned coverage of educational measurement regularly gave way to the immediacies: politics, storytelling, evaluation studies, conferences, reunions, and others. Short writings occasionally popped up, and, when not absorbed into some longer chapter or proposal, and sent elsewhere, became available. In 1974, the prospective Introduction I drafted went like this:

1974

I would like to report a theft. Someone has stolen the concept of educational measurement. I think it was before April 29, 1973. I looked for it -- and found it missing.

I had no trouble locating the concept of measurement of attributes of students. The cupboard of tests and test scores was certainly not bare. Many of my colleagues have found it easy to replace educational measurement with student trait measurement. Quite a technology has been created. But education is not measured by this technology.

It is true that educators are measuring lots of things besides the traits of students. Other people and other things are measured. Economic, physical, and attitudinal attributes are measured. Almost completely missing is the intention of measuring these things for the purpose of understanding educational situations and issues.

I am pretty sure I know the thieves. I didn't know the early ones, but later it was my coterie's books and papers on "educational measurements." They meant "tests."

But not always. Lee Cronbach,[1] *in his chapter in the second edition of* Educational Measurement *wrote: "For simplicity I refer to tests and test scores ... The statements, however, apply to all procedures for collecting data, including observations, questionnaires, ratings of artistic products, etc.*

[1] Cronbach, L. J., 1971. Test validation. In Robert Thorndike, editor, *Educational Measurement,* 443-507. Washington, DC: American Council on Education.

Most statements apply to protocols and qualitative summaries as well as to numerical scores."

One problem is that the same measurement statements do not apply to description of both particular and general situations, do not apply to explanations as well as to personal understandings, do not apply to the use of standardized and day-to-day observations. Our technical and evaluation language fits educations problems poorly. Our measurements invite few teachers, students, parents, officials to dig deeper into education's problems. Do the vast body of course dialogues help us understand teaching and learning? Do our measurements guide comprehension of critical moments in the classroom?

So by even those early days I was feeling doubtful. I felt we knew what should be measured but that we should be much more careful in getting the scores used right. It was only slowly that I saw that the scores we reported were not just misused but that we were compliant in accepting assignments and contracts that distracted from understanding education. Here's how I put it in 1972, an essay first placed chronologically in this book, but pulled back now to this Introduction. It starts with a quotation from Daniel Patrick Moynihan:[2]

1972

A century ago, the Swiss historian Jacob Burckhardt foresaw that ours would be the age of "the great simplifiers," and that the essence of tyranny was the denial of complexity.

[2] Moynihan, D. P., 1970. A quote from his farewell speech to the President's Cabinet in 1970.

He was right. This is the single greatest temptation of the time. It is the great corrupter, and must be resisted with purpose and with energy.

Moynihan's words are words to savor. There may be answers to many questions here. I wonder if there is an answer to the question "Why do I continue to be a measurements man?"

I am not a man of change, so part of the answer is "inertia." But I am a man of purpose, and I see that the long-range purposes usually given for educational measurement are seldom achieved. For example, sophisticated use of test results is not prominent in contemporary education even though in-service instruction in test use has been prominent in this country for sixty years. Worse than that, our formal measurements are not directly useful in the solution of most important educational problems.

But I have known *this* for a long time. "Why do I continue to be a measurements man?" The answer I usually hear -- it is an answer I have believed until recently -- is that measurement leads to analysis. Analysis leads to rational thinking. We measure in order "to tell it like it really is." When we know how it really is, then we can consider the alternatives and their implication and make the rational choice. Good choices will increase our control over our destinies and postpone our succumbing to them.

For myself and my fellow men, I want a freedom of choice, a chance to control our destinies. Science, technology, and measurement are the instruments of choice -- so that answer reads.

I am increasingly less persuaded by that answer. I do not see people becoming more rational even when their measurements are better. I do not see people increasingly in control of their destinies. Just the opposite. I see them more isolated from control and feeling more alienated by the increasing demands and constraints of commerce, transportation, war, protests, and government.

People do not see our data as key to the solution of their problems. My measurements are praised sometimes by my students and colleagues, seldom by real clients. Will that change?

Science and technology seem to be contributing less to our enlightenment, more to our alienation. Sometimes the client tells me point-blank that he doesn't want any more of my coefficients. But I say that he still guesses that the potential misuse of my information outweighs the potential good use. For the present he may be right. But my productive life is half over. Will he soon fear my measurements less?

The problem runs even deeper. What am I measuring? I have lost the sense that there is any "the way it is." What *is* only seems to be. “What seems to be” I can measure more accurately and reliably and verifiably. It does not mean that I am measuring what is. In fact, by de-emphasizing temporal impressions, clinical judgments -- those personal determinants of the world -- I am withdrawing from the challenge of measuring what is. What they see and feel is what is.

My measurements are not the first approximations to truth; they are choices I make as to how to clothe the truth. It is another case of the Emperor's clothes -- with the switch that it is the Emperor who is invisible and it is only the clothes that are seen.

Have my measurements no more purpose than to stimulate my fellow specialists and to delude the others! I think so. There is a purpose. It is a purpose seldom recognized, seldom honored. I think measurements help counter that onrush of the Great Simplification.

Philosophers and technologists are colleagues in the Great Simplification. Only observations stand in their way. Data occasionally support an idea, often not; seldom do they confirm an idea. Sometimes they, simplifiers, use research data to argue a point. But the principal effect of research -- as we see it today in education -- is to deny the validity of the hypothesis. Measurements always say, "No, it is not quite like that." Measurements are seeds of doubt.

The world needs advocates; the world needs skeptics. I see a world that scrunches its skeptics and counter-advocates, glorifies its advocates, each in his time. One truth, one set of values, one perspective, is too much and too often honored over others. I see measurements as vital to this world, not because they tell us what is truth but because they keep other sides of truth alive.

Bob Stake, 1972

Where's the evidence that measurement is keeping truth of education alive? In 2022, educational measurement is no closer to representing educational affairs. It would be wrong to exclude the notion of student testing, or census taking, or grading – for measurement can serve those ends. But fundamentally educational measurement should be both the miniscule *and* panoramic representation of all teaching and learning, all sharing of understanding. It would include the ambiance, the calendar, the politics, the sacrifices, the subordination, all consequent circumstances as well as the immediate matters of teacher remediation and student insight.

In a narrow sense, measurements are those discriminations, such as achievement scores and floor space totals, beyond purchase of the unaided eye; but in a larger sense, those recordings of education-at-work, particularly useful when there is little opportunity for direct personal observation.

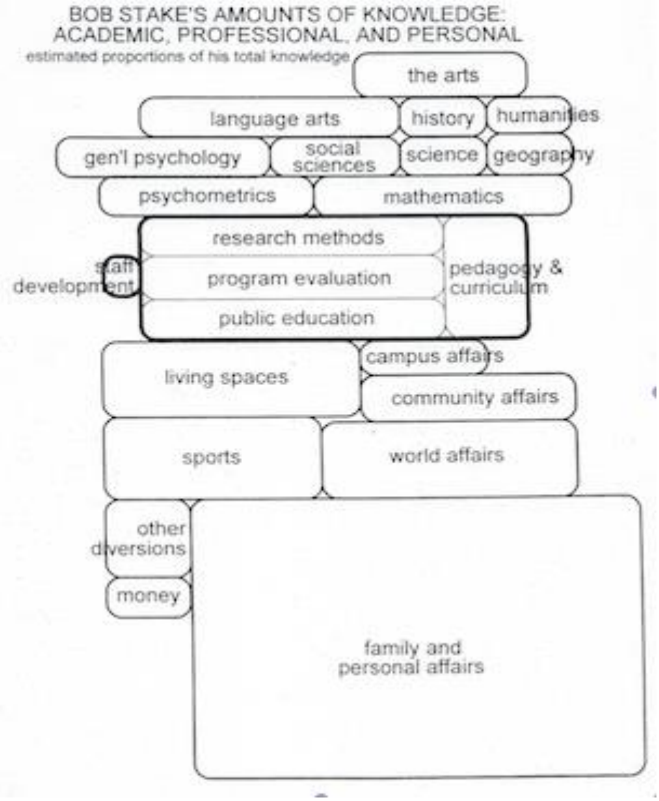
Here now, I am tying together loose pages for a book mostly for family and friends, but for you, _ _ _ _ _ _ _ _ especially. All of us are educators giving these topics a good piece of mind, as I have done these 58 Illinois years.

Bob Stake, 2022

1965

I joined the Illinois Faculty of Education in 1963. Shortly thereafter, Rupert Evans became Dean. And a little while later, he asked me if I could do anything to help him understand what the college as-a-whole was doing. Or maybe had done. I was pleased with the assignment, not thinking of it as spying, and went off to figure out how I might map the work of the faculty. Or maybe the knowledge of the faculty. I don't know what I might have needed money for, but the University gave me maybe $1000 for expenses. After perhaps a year, I gave the money back and told Rupert that I was making no progress. I realized that this was not just a Dean's need, that each teacher ought to have a way of visualizing the history of all students' engagements, at least an approximation of all their knowledge. So, still guessing, I mapped my own knowledge. I was happy with my rough map, but I still couldn't figure out how I might get such data from students and faculty members.

Mapping Knowledge



1967

This story had a bad ending and happy ending. I forgot and left the only copy of my next day's conference speech at the campground. So I winged it, and it came off fine.

Hybrid Seed Corn

Our tent was up, and we easily could have lasted a rainy night. But we went into town and found a motel. We had wanted to see a TV special that night so it was easy to leave the storm-threatened Pennsylvania campground.

The TV program was called "Leaving Home Blues." It was about young people leaving farms and farm communities for the city. We viewers were struck by the sight of an almost endless line of southern youngsters striding across the stage to pick up a diploma and striding right on to the Trailways station to buy a ticket to New York. The home-town ties in Nebraska and Texas appeared just as tenuous. The cities had few jobs for these youngsters, but hometowns had neither jobs, nor companions, nor hope, so they left.

The cause of the exodus was clear. There were no jobs because agricultural technology had succeeded. Tobacco could now be raised and harvested without a large supply of semi-skilled and unskilled labor. Corn was now most efficiently grown on the highly mechanized corporate farm. The one tractor family farm was a thing of the past.

The future for these young people is not clear. The problems of the city seem gigantic, much larger than the original small town and corn-production problems. City problems are aggravated by the influx of job seekers, yielding little to their faith and vitality. These are not now the cities' problems but the Nation's.

The shock of the TV program to me was the role of the land-grant colleges. The original script (I had been told) held industry, agri-business, and university research all at fault, but the land-grant college came off being the villain in the version we saw. The tobacco harvesters and hybrid seed corn were developed at experiment stations at the land-grant universities. It had not occurred to me before that the experiment station was responsible for much of that activity down at the bus station.

It was a shock because I had long looked at the experiment station with envy. As I saw it, researchers there had been charged with increasing productivity and reducing costs. They did the job. As an educational researcher I had been urged to increase the quantity and efficiency of learning in the schools. I had not found the way nor had other educational researchers. We could not point to a breakthrough close to that of hybrid corn.

Now, I do not know whether to be ashamed that I do not know a better way to teach kids mathematics or to be delighted that the comprehensive school has not suffered the fate of the family farm.

Of course, during the many years we have been seeking ways to systematize instruction and to standardize assessment, we have said that in a technologized school the teachers could be relieved of menial tasks. With our inventions to take care of routine teaching operations, they could concentrate on the talents, goals, and motivations of individual youngsters. The teachers would be free to improve personal inquiry and relationships among people in class and out.

I am no longer optimistic about that. I doubt if many teachers can and would pursue different goals than they now do. I do not see teachers being that flexible. And I doubt if the citizenry would continue to hire teachers of higher-thought processes and human sensitivities if they found that a technologized corporate school could teach the basic academic and vocational skills. A technological breakthrough in education is not to be feared because the wrong things would be taught but because too many of the right things would be shouldered aside.

It may be that arithmetic skills and music appreciation can be taught better by techniques developed by a regional laboratory and administered by a centralized traffic control station. Study halls, computer-aided instruction booths and school libraries might better be administered as public utilities. But what then would be offered as "the public-school curriculum" would probably be greatly different from the present curriculum. Prime objectives -- such as word recognition, knowledge of our ecosystem, knowledge of career alternatives -- would get increased priority. Secondary objectives (less immediate, more difficult to state operationally, more subject to controversy) such as how to work within a group, skills for evaluating evidence, working with humans having diverse values -- would get lower priority than they now have.

With technology these priority changes will not have citizens, philosophers, and curriculum specialists thoughtfully reviewing the alternatives and making the decisions. The technology would preclude the choosing. The nature of technical invention is to simplify alternatives, routinize operations, and bypass deliberation.

At this point the conclusion might seem to be that teaching technology should be sabotaged. But there is at least one more twist in the tape. To sabotage technology is to defy human design of all kinds. Much technology is essential. Much good is done even in the search for a better technique.

The line between machines and men is not the boundary line of technology. Technology includes any routinizing, standardizing, sticking-with-the-tried-and-true, quality-controlling activity. So that anyone who has a good way of doing something is a technician; a learner of his way is a technologist. To sabotage technology is to undercut experience, expertise, the promise of learning how to do something. Some technologies need to be sabotaged, but many need to be nourished.

The problem is more difficult because educators (and all folks) work more effectively in a system that promises reward for invention of new techniques. Better than they do in a system that promises constant skepticism. Skeptical reviews are important but curiosity and inventiveness and possible breakthroughs need to be encouraged. We should err on the side of enthusiasm. I think the human spirit requires it.

We should sabotage those technologies whose negative side effects outweigh their positive main effects. (Judgment is involved -- no universal agreement will be found.) The search for evidence of "negative" and "positive" should prosper -- though that search itself will have negative side effects. We should be cautious about switching to grand new ways of teaching without examining the social, economic, psychological, and moral costs.

Most educational technologists, just like seed corn researchers, are not experts in assessing and minimizing side effects and long-term consequences. Their job specifications, their research goals, will seldom be broad enough. Others -- professionals and non-professionals -- must watch, and protest, and pray.

We had no trouble getting back to the campground. Ours was a two-way road. Some others were stuck with life in the city.

1968

Increasingly I was aware that the academic disciplines were simultaneously protecting the precepts of the past and making up denotations for the future. It was not necessary for a test to be measuring something, just to correlate with something, and that was enough to be its "measure."

A is A

California recently opened the tailgate to let loose a bit of the frenetic, kinetic creativity of its people. Following the example of Massachusetts and New Hampshire, California has allowed each car owner a six-letter license-plate message. The best I had seen in New England was: *VANITY.* In Santa Monica recently I saw another mind-stopper: *A IS A.*

A is A. I thought for a moment of cryptic encodings, perhaps a poignant Alicia is Always, a patriotic America is Able, or a baseballish Anaheim is Angelic. But I couldn't avoid the conclusion that I was hearing a wee voice, a protest against an entwining, clover-leafed world, a world without verities or destinations, but profligate with passages and explanations.

A generation ago, in Broadway's *West Side Story,* the Jets gave Officer Crupke a lesson in explanation, pointing to the innumerable "causes" of their delinquency. Remember that? We were left with the notion that too much cause is no cause at all.

All the important things today seem to have too many causes. (Enlightenment, thy name is license.) There are as many explanations as there are explainers. Each explanation is a viable challenger, each a laser beam of persuasiveness, none a sunbeam of simple truth. Each expert has his say; the phenomenon is draped, garland on garland, often times hidden from view.

“Is this any way to run a railroad?” said the owner of the scarlet MG. He ordered a banner from Sacramento, planted the standard firmly in the passing lane of the San Diego Freeway and aimed a frozen eye at those who claim that A may not always be A.

Well, railroads aren’t run the way they used to be. Today each man is his own dispatcher. Adam is Authority, Albert is Authority, Alicia is Authority, ... Pope Paul! Abby Lane! George McGovern! Joe Garagiola! Sorry. One man, one vote. And "Sure, you can change your vote.” This nation is now founded with the proposition that all men are created equivocal.

If there is no one enduring Truth, is there no Truth? What is A if not A?

Unamuno, in his Castillian Spanish, said something[1] that -- on the margin of a student's notebook -- became: *Let us live in such a way, that, if there is no life hereafter, it will be a damn shame.* Whether or not there is an A transcending all notions of A is not important.

The important thing is to behold the many notions of A as we can best behold them. We cannot say that this and only this is A. A is living, growing, changing, splintering, splendoring, garland on garland. A, the eye of the fly, each eye of every fly.

[1] In *Representative Spanish Authors, 2,* Oxford Press, p 516. W. T. Pattison translated Unamuno: *And if what is reserved for us (after this life) is nothingness, let us act so that this will be an injustice.*

1969

This is more or less the speech I gave in Boulder, Colorado on October 20, 1967, a Friday before a Big 8 football game. I was expressing my concern about the shift of research funding from the production of research to the dissemination of research. It seemed easier to spread the word than to create it.

Information and the Ever-Normal Granary

When I visit Boulder, I am impressed by the sweep of 1and to the east; prairies rushing westward, unbroken except by the upward thrust of an occasional grain elevator. I am impressed by the prairie more than by the mountains, especially when I can climb into the mountains or these foothills to look out upon the prairie, now fields of grain, extending into my own Nebraska -- with its indomitable populism, and invincible football teams -- some of the time.

I've had those grain elevators in mind for a while now. With a shudder, I think back at drought-time dairymen dumping milk in the creek, farmers killing baby pigs, back at the ever-normal granary, food stamps, price supports, later the soil bank. Are we in education headed for similar scandals, the scandals of affluence in a world of need.

I've been worrying about the information explosion. I see us already information rich -- and headed for super-abundance. Hundreds of new agencies have been created in this country to get more and better information to practitioners, in our case, to educational administrators and teachers. The promise is that professional decisions will become more rational, overt, and public-minded; less intuitive, covert, and impulsive.

My home base at the University of Illinois is a service agency. We call it CIRCE. We promise to try to help. Our patrons are educators who need to evaluate some kind of program. We hope to help them with the plans, the ideas, the checks and balances, the information with which rational educators can do their job. We seldom subcontract to do the evaluation for them. We do not build instruments. We do not train their staffs. But we do have information based on experience; some of it abstract, some of it practical, some of it high-blown and some of it earthy, and they can do with it as they please.

We think our information is good, some of it better than you can get in Lincoln or Pittsburgh or Santa Monica. And it's different, I think, calculated to give some new ideas to even the seasoned researcher and administrator. Instead of emphasizing the research comparison of experimental and control groups (which is important for many kinds of research) we emphasize direct classroom observations by observers trained in different disciplines: e.g., anthropology, philosophy, sociology ... Instead of behavioral specification of objectives (which is important for much work in testing) we emphasize preferential scaling techniques to assign priorities to activities to observe. We advise our colleagues to gather full descriptions of what is going on and diverse opinions of its value. We offer advice, some good information.

We work hard with the notion that the purpose of formal evaluation is decision-making. Our evaluation plan starts not so much with a statement of instructional objectives as with a review of decisions made and decisions forthcoming. Once we have a good idea of what alternatives are facing the decision-maker, we or they can collect information on means-ends relationships. And we are likely to uncover a number of additional decisions -- previously unrecognized -- that need information if rational choices are to be made. Dan Stufflebeam at the Center for Evaluation at Ohio State takes a rather similar stand on evaluation plans. He has worked with Dave Clark and Egon Guba at Indiana to discover the contribution evaluation can make to innovation of schooling.

You can get advice more behavioral, more "Tylerian" and "Skinnerian," from Tuscon's Project Epic, from The Southwest Regional Laboratory in Los Angeles, and from the AAAS science-curriculum evaluation team. Their focus is on criterion-referenced behaviors, those particular student achievements specified in teaching objectives. From those places, Bob Hammond, Dick Schutz and Henry Walbesser are saying that we deceive ourselves if we conclude that a program is good because we have evidence that it was planned well, that it has commendable resources, that it is logical and relevant, that it is ethical. They want proof of the pudding in the eating -- they want to see changes in pupil behavior. So sometimes do we, but we don't trust such changes to tell how well the educators are doing their jobs.

I am not sure what kind of pudding they have in mind. Bread pudding? My brother-in-law Bert Evans, an agricultural economist at the University of Nebraska has been studying the U.S. bread industry.[1] He has reported that you cannot change the consumption of bread -- it's constant. "Inelastic," I believe he says. Many markets are susceptible to advertising and promotions, but not so much the bread market. If your football game broadcasts are sponsored by Sunbeam Bread, as ours are, it's because that bakery is protecting its share of the market, not because it is trying to stimulate the eating of bread.

Bert's advisor at Harvard was John Kenneth Galbraith. Galbraith has not been talking about bread lately, as Bert is. Galbraith has been talking about "the new industrial state," about the control big business and big government have over

[1] Evans, B. M. co-authored with Walsh, R.G., 1963. Economics of Change in Market Structure, Conduct, and Performance: The Baking Industry 1947-1958. *University of Nebraska Studies,* New Series 28.

the markets[2]. He takes it as already established that the demand for products is not determined by preferences of rational consumers in a free market but is determined in large part by advertising and manipulation and borrowing.

Among the people who are available to help me make up my mind what textbooks to use or what computers to buy are field representatives of IBM and Rand McNally. They are, I feel, do-gooders. They help me find out various things. They are candid, straight-shooters, for all I can tell. So far, anyway they have worked at keeping my trust rather than making a sale. I look for a lot more contacts to be made with teachers and administrations in the future.

Have you heard of EPIE? That is the Educational Products Information Exchange,[3] located in New York City. They are good guys too. They want curriculum specialists and purchasing agents to have access to information about instructional products. They want to provide information to users of textbooks, films, language labs, etc. They want to enable buyers to know all the choices and all the grounds for making a choice. They want producers to worry more about pilot runs and technical manuals, and worry more about claims in advertising, and worry more about consumer needs still not met. Through texts and charts, EPIE hopes to be in touch with the classroom teacher.

A fellow telephoned Terry Denny at EPIE this week and said, "We don't need you. Litton Industries is doing every field test you are planning to do." That was news to Terry. But there are lots of helpers already. The State Department of Education of Pennsylvania is farther along analyzing mathematics textbook information than EPIE. The ERIE Regional Lab has similar aims. Many Title III Supplementary Centers have a

[2] Galbraith, J. K., 1967. *The New Industrial State.* Houghton Mifflin.

[3] Euchner, C., 1983. Kenneth Komoski Helps Wary ‘Consumers’ by Evaluating Computer Products for Schools. *Education Week*, February 2.

Consumers Report service like this in mind. Many state departments of education plan to help.

I looked over my copy of the Summer, 1967, issue of the AERA newsletter, the *Educational Researcher*. It was a special issue on information systems, storage and retrieval. More than a dozen new educational data banks were identified. Most of them intended to provide the driver in the driver' seat with a better map, a better destination, a more detailed list of campgrounds and comfort stations. The equivalent of.

For the past two years or so the U. S. Office of Education -- although it is anything but a creature of one mind and purpose – has arranged much of its spending on the premise that there is an obstruction in the flow of information from researcher to practitioner, and that the practitioner would teach better, innovate more, and cure more ailments, if more and better information were available. To my mind, this is an unwarranted premise. I see already an abundance of information. School leader needs some of it, they know they needs it, but somehow it is out of reach. Perhaps we have not provided the information in usable format. Or perhaps it is not available at the right time. Or perhaps we are wrong in expecting or even wanting him to be more rational, less intuitive. Perhaps the information should be reconstituted as some form of experience for the experientially-oriented intuitive educator.

Back in the Thirties we became more aware that some years we could produce more food than we could eat. Not more food than the world could eat, but more than we could get eaten. A number of things were tried. One was the ever-normal granary. Following Isaac’s son Joseph's lead, we should produce extra during the years of plenty. "Lean years” would come. Some years we raised more than we could use. We gave some to the poor and some to the hot lunch program and we found other ways to use the surplus. Still, the concept of ever-normal granary didn't solve the problem. Of hunger or understanding. The embarrassment would be still greater if we had not found ways not to grow so much. The world is not well-fed today, but we have trimmed production somewhat to match consumption.

I do not like this as a solution for educational data surpluses. I do not want to discourage the production of information. A Nebraska bumper sticker says, "Eat more beef." Should we have bumper stickers that say, "Use more data." "Plan ahead." How about “the Faculty that *perts* together *spurts* together?" Is data obscurity correctable by advertising? I think not. I will be explicit, finally, about what is NOT my concern. I am not concerned about surpluses of information. I am not concerned that we have too many research projects, too many surveys, too many inquiries. These efforts, as I see it, are seldom capricious. They happen because somebody is trying to find out something. We should tolerate a data overload if it may help schools improve.

I AM concerned that we are spending too much for institutionalizing data systems to collect, store, and make retrievable educational research findings. I am concerned about the investment. We do not have sufficient reason to believe that practitioners will use what we produce. Information specialists will say, "How do we know until we try?" They say," How do we know what people will use until they have the chance?"

In the movie, *Field of Dreams*, Kevin Costner said about a ball field in a corn field, "If you build it, they will come." Context is as important as content. Not everything built will generate a field of users.

I sometimes go along with that argument: "Yes, build it." But now I am reluctant. Money that could be spent on research and training is today spent on the information dissemination buildup. Funding seems easy to get. Proposals are based on dreams, not on good marketing studies. For the moment at least, I need better evidence that we can count on practitioners being rational information users. If that is too much to expect, then we are spending too much money to provide it.

But here's a different thought. Maybe we should not want to draw Sandy Davis or Mr. Novak or Superintendent Smith more into being rational, research-based decision-makers. To what advantage? At what cost?

Kenneth Boulding[4] -- the outstanding Colorado economist -- in a book called, "The Impact of the Social Sciences," mentions three kinds of human knowledge: folk knowledge (personal experience), literary knowledge (the writing of experts) and scientific knowledge (information gathered by observation and testing). All three can be useful to us. Boulding suggests that scientific knowledge is not always the better knowledge. He suggests its inappropriateness for teaching how to behave at a sporting event or how to find the post office. Some matters, including professional matters, fit nicely into the domain of personal experience, others do not. Managing an economic system or planning an astronaut's journey requires a preponderance of scientific knowledge. It may be, however, that most of what a teacher does in the classroom, what most teachers do, should remain under the control of folk knowledge, intuition, and personal experience. If so, we should hesitate to encourage shipments from data banks to teachers. If so, we should not presume instruction will improve if more and better research information is disseminated.

But I have faith that, even then, better information can make a difference. Even in a situation dominated by highly intuitive teachers, instruction can benefit from the educational sciences. Change should not be dependent on a teacher's ability to digest technical information. By offering teaching aids,

[4] Boulding, K. E., 1966. *The Impact of the Social Sciences*. Rutgers University Press.

illustrations, opportunity to experience adapted to the technical data, we should be able to articulate modifications in practice. We should become better able to coach, guide, and shape teaching behavior -- at least to get new ideas tried out -- without institutionalization of the distribution of research findings.

I have been impressed by the optimism and style of the PLATO computer-assisted-instruction project at the University of Illinois[5]. Don Bitzer and his team expect to change professors into computer teaching users by engineering alone. They say we don't need professors to come learn our language, our knowledge, and our system. We are going to engineer the professors' lesson writing station to conform to the advanced lesson-writing station. He won't need to realize that he is now working on PLATO instead of his blackboard or typewriter. Gradually the bonuses of the computer system will come to his attention, and he will take advantage of them. With proper engineering (which the programmed instruction movement did not get, they say) the professor need not participate in the information sciences to be a beneficiary of them.

How can the PLATO people dare to have so much faith in their engineers? Should we demand a little more from ours? Can we insist on assimilation, as well as collection, storage, retrieval and dissemination?

Manfred Kohen[6] while working at the IBM Research Center, described their approach to information science. It emphasizes: "(1) continual adaptation to the relevant real world, (2) a division of labor between human and automatic elements of an information system, with effective communication between them, and (3) methods for automatically representing the relevant real world (for mapping it) to guide decision

[5] Dear, B., 2017. *The Friendly Orange Glow*. Pantheon.

[6] Kochen, M., 1984, Information Science Research. The search for the nature of information. *J. Amer. Soc. Inform. Sci. 35:*194-9.

making." It seems to me that we can only justify our present growth of information banks and services when we have shown we can transform the information to practitioner meanings. There is a real world of bread and granaries.

1971

Today some of us remember the hippies and the Viet Nam war, but have forgot Lyndon Johnson's war on poverty. Because of Sputnik, we put a lot of effort into improving the school curriculum to upgrade the brainpower of future leaders. By 1971 we were no longer supporting bringing education to the farmers with the university's county agents -- who were not as up-to-date as the leading farmers locally. But we knew that something should be done to help those having trouble in school. Brother-in-law Bert Evans had an idea.

Helping The Disadvantaged

Perhaps we at CIRCE should propose research and development on community groups trying to get educational reform to aid the disadvantaged. Teacher qualification development could be one element, but there would be others. One possible way was suggested in an interview with Bert Evans, a specialist in rural economic redevelopment at the College of Agriculture, University of Nebraska. Bert proposed a new land-grant university, field oriented, bootstrap operation, to get local people to study their local situation to help people with disability. They would hear about issues and discuss them in workshops. The Universities would provide program coordinators, local organizations could arrange the meetings, and the federal government would provide some materials and resource people.

Bert indicated that few rural citizens are alarmed about the quality of schools or the inequities of the poor. They will come to community meetings on tax problems, school consolidation, crop and livestock protection, and even community economic redevelopment. They will tackle tough educational issues at meetings called for other purposes, but there is little impetus today for community action to improve schools to serve the disadvantaged in rural areas.

Bert knows that the problems of education for the rural poor are large but that neither the poor nor the larger community groups are ready to address them. He doubts that they will -- except to protest school happenings they don't like -- until they become aware of the integral way schools are related to other systemic problems: taxes, roads, subsidies, jobs, loans, welfare, etc.

There is no question in his mind about the poor quality of education in these communities. The aggressive members of the community get their intellectual stimulation from outside the community and let the schools be custodians of ritual, mores, communication skills and basic facts (in a historical and academic sense, but not a social sense). The schools are very conservative and most in the community are content with it that way.

Bert feels that the Universities contributed greatly to the problem when they assumed that their traditional academic programs should work at providing the basic knowledge needed and concentrated on educating an already-advantaged pre-vocational student body. They should have been more sensitive to the knowledge-explosion impacting ordinary citizens and their communities. Such people are out of touch with the knowledge-side of the university. So they vote against school bonds and they cuss hippies (who in fact share many value-positions with them). The Universities should have almost all students and faculties off the campus and in the field most of the time, Bert said, not because theoretical problems can be found better there than in textbooks, but practical problems can.

At least in most social science problem areas the motto should be, "Help communities to help themselves." The

Universities (and governments) are not going to know the answers to those problems. They will know some things that community members do not, and they should know something of how to get community members to readdress themselves to these problems. Community redevelopment specialists, Bert says, have found this to be a relatively simple skill. (What better way to give the federal government a fighting chance than to help them with at least one solvable problem?). We have here in "greater CIRCE" a stronghold of community-oriented people: Tom Hastings, Jack Easley, Gordon Hoke, Mary Ann Bunda, Trey Coleman, Sally Pancrazio, Dennis Gooler, Terry Denny, Arden Grotelueschen, Hubert Dyasi, Ernie House, ... Shouldn't we put our evaluation ideas to work on Bert's commitment to the disadvantaged?

1972

Invited by Gary Joselyn, I presented a version of this paper at the 22nd Annual Conference on Minnesota Statewide Testing Programs, Minneapolis, September 21, 1972. I expressed my opposition to current efforts to move school curricula behaviorally and toward mastery of "critical learnings" and corresponding achievement testing. Back then I made the title, Off the Critical List*. "Off" was being used as a verb for "Down with ..." as in "Off the War!" By 1973 I had changed the title to "Beef Stew."*

Few learnings are critical. No performance objectives are suitable for all learners. Even though not critical, some learnings can be said to be prerequisite because it is convenient to have all learners learn them together. Mastery learning is wasteful of learning opportunity.

Beef Stew

Many people want us to teach: "the basics." Some school persons, especially curriculum developers, are spending large amounts of time identifying "critical" learnings. Many states have set the same basic scholastic goals for every child in the class. Lists of performance objectives or critical learnings are created as an aid to education -- but they can be "misleading." Such lists hurt as much as they help.

I refer to any list or device we use in charting student progress. It can be a list of needed knowledges or skills or attitudes. Any desired student accomplishment could be on the list. We call it a *critical list* if it is what somebody thinks "has to be taught everybody."

Such a recipe is the substantive base for any educational achievement test. And even a so-called intelligence test is a test of previous learnings. The list often appeared in the form of a “content by process” grid and an item pool. Standardized tests are built from grids and item pools -- and expected by some people to reflect what *every* student should know. Teacher-made tests are usually built from implicit lists -- and are expected by some to reflect what every student in the class should know.

Are any learnings essential?

If the question is, “Are there some things that some people would like everyone to know?” the answer of course is yes. As a matter of personal preference, anyone may declare some learnings critical. And public and professional preferences aggregate to something important in curriculum building.

If the question is, “Is there some competence that everyone must have?” the answer is no. Much is hoped for, but the world is neither surprised nor untracked by low competence. What is deemed "satisfactory" varies too much with the situation. Success can be attained in many, many ways. For schooling prior to professional and vocational training, only broad categories of thinking and only the broad curricular areas can be thought likely to be helpful. There is no one recipe for beef stew.

Within a culture there is a positive correlation between literacy and economic success, but some illiterates get rich. Many not very sophisticated people lead happy, respected lives. Perhaps the question should be, “Are there some learnings needed in order to keep alive the availability of an important subsequent learning?” The answer then is yes. For any course of study, an applicant is or is not "admissible" to higher learning. In any particular course of study, there may be prerequisites.

Doors are closed to persons without a "proper" background. Part of the grounds for admission to advanced schooling, to jobs and to privilege is evidence of eligibility for further schooling. The learner and his caregivers will treat many learnings as such evidence. It is in the culture. The "system" demands it.

The educator too is a victim of the system, but also an instrument of it. The teacher can make it more accommodating or less. Educators may say, "You must do these homework problems," or "You must have two years credit in a foreign language." If they say so, even partly with authority, they do so as an article of faith. They will not have evidence that those studies enable a person to learn more or to perform better than he or she would have otherwise. The point is that the educator should seldom imply that requirements are in the best interests of the student, except in the sense of "admissibility." School counseling expert Ralph Berdie has said,

> *... nothing good in and of itself is to be found in learning first year college mathematics; but learning first year mathematics allows a student to move into second year mathematics, However, nothing of ultimate value resides in learning this, but learning second year mathematics allows a student to move into further mathematics or to behave differently or learn about related disciplines, for example, physics. In turn, nothing about physics gives it ultimate value, but learning about physics in turn leads to other kinds of behavior, other experiences, none of which in and of themselves have ultimate value except insofar as they lead to further experiences."*[1]

Algebra is taught before calculus. By rule, most Minnesota children study Minnesota history. I do not object to

[1] Berdie, R. F.,1968. If there's no heaven -- On purpose and experience in education. *National Catholic Guidance Conference Journal*, 12, 4.

required subjects or prerequisites. We should remember that we have no formal evidence that an awareness of Minnesota history makes us better students, or better Minnesotans. We should teach it of course if we desire having youngsters exposed to state history. That is reason enough.

As to sequencing: Any evidence or logical persuasion that algebra must be learned prior to calculus holds only for some ways of teaching calculus. As a convenience or facilitation to the teachers of calculus we may justifiably call for prerequisites. These conveniences make instruction more manageable. But we are not justified in identifying such studies as *critical* prerequisites for learning, except in the particular context (e.g., with a certain book or teacher) they have been found to be sequentially advantageous. There is no general research base for criticality of sequence of curricular content.

Research on learning hierarchies has been confounded, I believe, by a failure to keep in mind the distinction between an acquisition hierarchy and a constituency hierarchy. An acquisition hierarchy identifies a critical sequence, i.e., acquisitions that are essential prior to final learning. If students *can only* understand "supply and demand" if they have previously learned the meaning of scarcity; or if students *can only* learn multiplication if they have a prior understanding of addition; then we have an acquisition hierarchy.[2]

The constituency hierarchy is a name for an increasingly-complex mixture, e.g., beef broth; vegetable soup; vegetable beef stew. An analysis may show that vegetable soup is consistently the combination of vegetables and beef broth, and that vegetable beef stew is consistently the combination of all the ingredients of vegetable soup plus beef particles. Then it

[2] It may be only a trivial hierarchy if it is just as efficient to learn all steps in a single instructional experience. It is important for instructional specialists to ascertain that they are working with non-trivial hierarchies before insisting on a particular learning sequence.

could be said that beef broth is a critical ingredient of vegetable beef stew, but not that beef broth must be acquired first. A constituency hierarchy does not reveal anything about optimum processing, only about ingredients.

Nor does a constituency analysis even identify the *desirable* ingredients. Some of the constituent parts in a savory mix may not contribute to its quality. The fact that they are regularly found does not mean that they are to be valued. Not all the curricular content or mental skills present in successful learning can be considered desirable or critical.

The analyses of the *Bloom Taxonomy* (cognitive domain)[3] by Russell Kropp and Howard Stoker and their colleagues[4] are examples of constituency analysis. The ingredients are "lower" and "higher" mental processes. Their conclusion was that the claim for an acquisition hierarchy was weak. However, if they had -- with their regression analyses and factor analyses -- found a simple, substantive, unique additive at each level of the *Taxonomy,* it still wouldn’t mean -- what many readers of the *Taxonomy* infer -- that you need to teach information before you teach reasoning. A particular learning experience may make a student more "admissible" to a later learning opportunity, but often not more ready for it, and certainly not more deserving of it.

Why do we have critical lists (standards)?

List-making is a human compulsion. Specificity and order *do* aid comprehension and communication, so lists have

[3] Bloom, B. S., 1956. *Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain.* Addison-Wesley Publishing Company.

[4] Kropp, R. P., Stoker, H. W., & Bashaw, W. L., 1966. The construction and validation of tests of the cognitive process as described in the taxonomy of educational objectives. Cooperative Research Project #2117. Tallahassee: Florida State University, February.

gained a high status with those of us who teach. Each verbal entry may but tenuously represent its constituent, and the meaningfulness of ordering may be obscure, but the list can still be a valuable mnemonic. So we make lists, and to advertise the truth therein we sometimes add "basic" or "critical" to the title.

The critical lists I am talking about are formed partly in response to those who advocate a more analytic approach to instruction and who advocate equating program quality with student-performance quality. When we hear an expert say that programs should emphasize the mastery of skills or that teachers should emphasize the acquisition of basic knowledge, we are likely to believe that we are being told that *there are* specific knowledges and skills that all students should know. Think about what some of the experts on instruction and testing say:

§ Ben Bloom, with his emphasis on mastery learning;

§ Bob Ebel, with his emphasis on the knowledge purpose of schooling;

§ Bob Gagné, with his emphasis on hierarchies of learning.

Psychometrician Ben Bloom[5] has urged us to think of learning as the mastery of tasks. He has developed taxonomies of objectives and content behavior grids for improving the techniques of instruction and testing.

Measurements specialist Bob Ebel[6] has argued persuasively that the main purpose of schooling is to foster the learning of useful knowledge. He says they can do this job well, that they perform other jobs poorly, so that accountability should be focused on knowledge generation.

[5] Bloom, B. S., 1980. *All Our Children Learning.* New York: McGraw-Hill.

[6] Ebel, R. L., 1971. When information becomes knowledge. *Science, 171*, 130.

Learning specialist Bob Gagné[7] has recognized and promoted the logic of breaking complex learning tasks into simpler components and teaching these components hierarchically. Many national projects to develop curricula in the last ten years have followed his model.

These are respected men, competent men. They have recognized weaknesses in traditional classroom instruction, and carefully devised remedies. They offer rational solutions to real problems -- hoping people will follow their advice with reason and temperance. But people (professors and teachers as well as the masses) take their words, and other pleas for specificity and order, too literally. People responsible for curricula devise projects in which teachers and others spend hours and hours writing objectives, sorting test items, and listing facts to be taught. It is time not well spent.

Teachers and curriculum developers who pay close attention to Gagné, Ebel, and Bloom seem willing to forego the high-complexity goals in education in order to fulfill the more specific objectives of rote learning. They say, "We want both," but when you look at the objectives they define, the syllabi they write, and the tests they give, it is clear that they are not emphasizing experience with complex problems.

I do not know of empirical research that preferentially supports this specific knowledge approach, the performance-based approach, the critical list approach, to curriculum development. I do know a small amount of empirical research that raises doubts about them. For example, Illinois doctoral student Don Bosshart[8] got four experienced test authors to draft a test covering a chapter from a high school chemistry book. The two authors given a content-process grid to assist them

[7] Gagné, R. M., 1972. Domains of learning. *Interchange, 3,* 1-8.

[8] Bosshart, D., 1972. Evaluating instructional content. Urbana: University of Illinois, Unpublished doctoral dissertation.

wrote poorer tests on the chapter's main objectives than the two authors who did not have such a grid.

Instructional evaluator John Zahorik[9] found that teachers teaching from specific objectives were less willing to answer questions and utilize problems that the students brought forth. Learning researchers Lauren Resnick, A. W. Siegel and Esther Kresh[10] attempted to validate the hierarchial nature of a Gagné task but concluded that the learning could be achieved in various sequences.

These studies suggest to me that the notion of critical learning can be at cross purposes to the development of good curricula. My observations as a curriculum evaluator have persuaded me that critical lists can, in fact, impede the maturation of student minds.

What does a person really have to know?

A person has to learn some specific things, e.g., the simple meaning of zero, to read the word *Poison*, to feed oneself, to pull to the curb when the lights behind are flashing-red, and to smile, if on the most gorgeous spring day, someone says, "I could take a week of this, if it would then turn nice." But these specifics are not dependent on the help of the school.[11]

Yes, a person needs some language and numerosity, but little in particular. A person needs great networks of knowledge. Much of it must be complex, a basis for analysis and application, as Bob Ebel says. No one knows which networks will be useful

[9] Zahorik, J. A., 1970. The effect of planning on teaching. *The Elementary School Journal,* December, 143-151.

[10] Resnick, L., Siegel, A. W. and Kresh, E., 1979. The equivalence of positive and negative methods of validating a learning hierarchy. *Contemporary educational psychology,4, 3,253-259*

[11] Some things should be learned early rather than late, e.g., swimming and speaking without an accent, it if is important to learn them at all.

to herself or himself. Fortunately, experience shows that youngsters do acquire a useful and organic body of knowledge when exposed to parents, peers, and teachers who challenge their minds with interesting problems and complex human experience. The content and constituent skills seldom seem critical, but challenge and complexity do.

Again, Ralph Berdie: *Living consists of responding to experiences, and education consists not only of preparing students to respond to experiences, but even more importantly, of providing situations in which experiences can occur.*[12]

And experience shows that teachers, parents, and peers do not have to have well conceptualized philosophies and well-stated teaching plans in order to arrange good learning experiences. If other things are working, the grand plan will take care of itself.

Ben Bloom, Bob Ebel and Bob Gagné do not claim that certain learning tasks are critical for success. I have mentioned their names in this assault upon the critical list because educators who stress critical accomplishments invoke their techniques and their authority. But more: mastery-learning, knowledge-rich curricula, and task-analyses for curriculum building do not specifically rule out objectives any person might prefer -- but they do, in effect, draw the concern of educators and citizens toward the more-easily-stated and the more-easily-tested objectives. A spotlight is more effective if the rest of the room is dark.

This is the danger: an unwarranted changing of priorities. I believe that learning hierarchies and mastery learning almost hypnotize people into thinking that certain objectives, certain tasks, and certain test items arc critical ingredients in education. I believe that critical lists are bad

[12] Berdie, R. F., 1968. If there’s no heaven -- On purpose and experience in education. *National Catholic Guidance Conference Journal*, 12, 4.

because they lure us toward too great a focus, toward too much redundancy of teaching, and toward learning that is most likely to be tested.

Our present mode of life, heterogeneous as it is, does not require a standardized scholastic preparation. In the life that you and I know, the jobs and the problems seldom demand immediate, unaided response. We usually have time to ponder. We usually have resources -- other people or documents -- to help us. We often can ignore the problem without jeopardizing our standing. Life may be a bummer, but it is seldom a closed-book test.

Most teachers, most educational leaders, most people, have very little insight into true criticality, the causal relationships between school achievement and life success. There are lots of ways to be successful. Oftentimes a job stretches itself out to fit the worker. When we examine the life stories of successful men and women, we are impressed with the ingenuity with which those from diverse backgrounds overcome adversity. The calculated "correlation" between knowledge and common measures of success is high and positive. But the evidence of betterment-of-an-individual attributable to learning-any-particular-thing is information teachers do not have, nor that researchers have been producing. For good reason: the evidence would surely be transient, not generalizable, no longer relevant even to the place from which it was taken. We lack good reason to try to build critical lists validated against success in life.

If this were a time for mobilization of the national working force; if this were a time when everyone should be maximally productive; if workers were immobile, and jobs had fixed specifications -- we might be justified in a search for critical preparation. But it isn't and we aren't. The working force could work half as long and still produce more in hard goods, services, and ideas than we need for comfortable living. It is not a time for task analysis and mastery testing. It is a time to look after the quality of the beef stew.

1973

Written during the family's sabbatical stay in Sweden at the Institute for Educational Research, University of Göteborg, hosted by Urban and Tordis Dahlöff and Ulf and Gunbritt Lundgren; perhaps marking my drift from Psychology to Education. I thought of this at the time as a possible opening to a book on the Measurement of Education.

Learning is the original and natural state of human activity. Education is the product of learning, not vice versa. School is a place where youngsters learn how to deal with ideas and people. Teachers create settings more than learnings.

Genesis

In the beginning there was learning. No teachers were there. No bells rang that learning might begin. The natural state of the child was "learner." It still is.

Teachers came later. Schools came later. State departments of education, professors of educational philosophy, and the data-processing planning committee, came later. Learning was not indifferent to new creatures about it -- but it did not become *their* creation.

Even that first day Jacob went to school educated. Neatly combed outside, an intricate jumble of experiences inside. His mother's caresses, a bird and butterfly mobile, voices and faces, a chase down dark alleys, popsicles, and Sesame Street -- still there.

He came home, again and again, partially re-educated. New experiences added on, old ones turned about. New moves, new connections -- some expected, a few intended. Day by day, moment by moment: attention coming, going, in and out of focus. Dewey[1] said:

> *I assume that amid all uncertainties there is one permanent frame of reference, namely the organic connection between education and personal experience.*

The slate is never clean. The pretest score is never truly zero. The teacher is no sculptor, chipping formless rock, shaping raw clay. The child, even the shiest of the shy, is a stretching, bursting, consuming person, already formed, forming anew, seldom waiting choice-of-book or word-of-reinforcement. Earlier I expressed it this way:

> *Learning is a mountain stream, fed by untraced snow and rain, moved by every contour of its path, susceptible to diversion and dam -- if outside powers choose so great an incursion -- but in the usual course, drawn by unconscious appetite and aspiration, through adversity and satisfaction of the plunge, to destinations still belonging to the mountain.*

Jacob's teacher is a part of the environment of school learning, not a creator of learning, but doubly important as enduring presence and arranger of environments. Protection, mild constraint, immediate direction, solace, scenery; she provides much, she produces little. Almost never does she arrange or produce a different destination. The stream belongs to its watershed.

Society, the watershed, the creation, is ever changing, but yielding little to planned change and liberation. The schools are one extended-reach of society, an expression of its changes but not the instrument of its reform. Children learn the public

[1] Dewey, J., 1938. *Experience and Education*. New York: Macmillan.

truths[2]. Societies change, but they are not managed. Children learn, learnings change, and societies are what children become.

Paulo Freire[3] wrote that there is no neutral education: that education works either for domestication or for freedom. In terms of society's business, I see education only as domesticating. It sensitizes the child to the constraints of law and custom, passion and reason. In this world it is not possible for the schools to be the agent of freedom, to allow more choosing the ways of society than the society would offer. As education helps the child become more aware of society, it domesticates. And most people, even most of the oppressed of Freire's Third World, would have it so.

Educators themselves are not modest. They exalt the business of teaching, of schooling, in words such as these of William Cory[4], a 19th Century Master at Eton:

> *You go to a school at the age of twelve or thirteen: and for the next four or five years you are engaged not so much in acquiring knowledge as in making mental efforts under criticism. A certain amount of knowledge you can indeed with average faculties acquire so as to retain; nor need you regret the hours spent on much that is forgotten, for the shadow of past knowledge at least protects you from many illusions. But you go to a great school not for knowledge so much as for arts and habits; for the habit of attention, for the art of expression, for the art of entering quickly into another person's thoughts; for the habit of submitting to censure and refutation, for the art of indicating assent and dissent in graduated terms, for the habit of regarding minute points of accuracy, for taste, for discrimination, for mental courage, and for mental soberness.*

[2] There is no ultimate truth that education might serve, were it even so inclined.

[3] Freire, P. R., 1996. *Pedagogy of Freedom: Ethics, Democracy, and Civic Courage.* Lanham, MD: Rowman & Littlefield.

[4] Cory, W. J., 1861. *Ionica*. Smith, Elder & Co.

Children do acquire knowledges, arts and habits, in some measure, and with some help from the schools. Cory's conceit was extravagant, but I am pleased when my children are taught by such a teacher. Not that I believe he can make Jeff, Ben, Sara, and Jacob what they will not otherwise be, but because I want them to know human beings who believe in these arts and who strive to develop those habits.

The school is a liberating place, even if not a liberating force. It is a place for personal opportunity. Not only does it house many of a child's destiny of learnings, it houses: the printed word of past experience, peer learners from in and outside the neighborhood, and a frothing collection of adult mannerisms. These are rich surroundings for the learning child. He learns how better to deal with authority and authoritarianism, with ambiguity and specialization, with affection and affectation. He learns more about escape, whether in the reveries of the library, in the explanations of the laboratory, or in the seclusion of the basement by the boiler room. In a personal sense, the child is liberated by these experiences, unintended though it may be, and perhaps because of them he has a chance to be more what he would choose to be.

To provide these opportunities for experience is the precious responsibility of the school. Of course, John Dewey[5] was right when he said that not all experiences are equally educative:

> *Experience and education cannot be directly equated to each other. For some experiences are mis-educative. Any experience is mis-educative that has the effect of arresting or distorting the growth of further experience. An experience may*

[5] Dewey, J., 1938. *Experience and Education*. New York: Macmillan.

> *be such as to engender callousness; it may produce lack of sensitivity and of responsiveness. Then the possibilities of having richer experience in the future are restricted. Again, a given experience may increase a person's automatic skill in a particular direction and yet tend to land him in a groove or rut; the effect again is to narrow the field of further experience. An experience may be immediately enjoyable and yet promote the formation of a slack and careless attitude; this attitude then operates to modify the quality of subsequent experiences so as to prevent a person from getting out of them what they have to give. Again, experiences may be so disconnected from one another that, while each is agreeable or even exciting in itself, they are not linked cumulatively to one another. Energy is then dissipated and a person becomes scatterbrained.*

Dewey was later more explicit about criteria for desirable educative experiences (humaneness, continuity from one to the other, interaction between external reality and internal sensitivity), and spoke repeatedly about our need for a philosophy of education based upon a philosophy of experience.

I am sympathetic to this experiential conceptualization of schooling. But I recognize here too the conceit of the educator.[6] What makes one scatterbrained, slack and careless, or callous, is not a classroom experience, nor all the experiences a teacher can devise. School has no such power to create or

[6] It is the conceit of a God who tells Adam not to eat fruit from the tree of knowledge (Genesis 3) and the obtuseness of a Noah who keeps his entourage aboard the ark 57 days after settling on dry land (Genesis 8). The reason progressive education was rejected was more that people did not like what those teachers and students were doing, not that teachers failed to teach. Dewey was right about our need for well-conceived experiences without being quite right about the mis-educative effects of a given experience, and without being right about our need for explicit philosophy.

"taketh away." If I thought it did, I might not let my Jacob go near a school.

I do want him to share in as much of "the Good life" as he can. While he is young and all his life, the Good Life includes learning, and it includes school, for all the shortcomings. Curriculum and teaching method are important. They do not determine what way nor how far Jacob or any other child will go. But each fashions better or poorer opportunity to know the world, to increase the pleasures of contacts with others; and to ward off their encroachments. I do not expect Jacob's teacher to know when those opportunities will occur, and I especially do not want her to plan in detail for them. I want her to arrange experiences -- humane, with continuity and interaction. To do this she, the teacher, and the principal, and the Commissioner of Education, need to have a sense of what is happening. I want them to know *that* more than they do.

That is what this book is mostly about -- how measuring things ought to be helping educators know the power, the futility, and the peace-that-passeth-understanding of teaching, learning. Learning is the original and natural state of human activity. Education is the product of learning, not vice versa. School is the place where youngsters learn how to deal with ideas and people. Teachers do not create learnings, they create settings.

1973

Here is one of the essays I wrote on sabbatical in Sweden, visibly a land of humane treatment and civil liberties. You may feel that not much was rubbing off on me. Galton was probably in my extended family tree and I expected him to be one of my role models. Other mentors, Harold Gullickson and Warren Baller, spoke favorably of the Eugenics movement. Arthur Jensen and Charles Murray were to become acquaintances and, back in time, I didn't yet take offence at their racial views.

If I were to be a psychologist, it would be within "individual psychology," working on how the different propensities of individual students could be developed in school. I took my job at Illinois in 1963 to bring testing and teaching closer together. As appears in my 2018 essay, only gradually did I see the hurt of "everyone ever-lastingly tested."

Galton's work became the discipline of psychometrics, partly through the works of Lewis Terman and Leon Thurstone, mentor of Harold Gullickson, my academic father at Princeton. Thurstone's research shaped my dissertation,[7] *"*Learning parameters, aptitudes, and achievements*," Not long after, I realized that I was not ready and wanting a career in mathematical psychology. Years later, I realized that Galton and Thurstone's work on psychometric scaling could be seen as precursor to my ultimate emphasis on "experience" as fundamental qualitative-ground for differentiating among humans. And still later, as I say in my 2018 essay, "Those Not*

[7] Stake, R. E., 1958. Learning parameters, aptitudes, and achievements, (doctoral dissertation), Princeton University.

Chosen," I came face to face with the dark side of discrimination, even by educators of high purpose.

An Anthropometric Station

On the south side of Eno Library at Princeton was something of a shrine. A small wall there was the only one not covered with photographs of elder but-less distinguished psychologists. There were but two nails on that south wall, one above the other. Every once in a while, we psychometrics students would have to go in and put Galton's picture back above Freud's.

Francis Galton was the father of individual psychology. More perhaps than anyone else he systematized the search for ways in which persons differ. He believed in -- in contrast to Freud -- the physical determinants of behavior. In 1884 he created an "Anthropometric Station" in South Kensington, London, to measure the physical and psychological powers of London school children and adults, at their expense, three-pence. Some saw his views as assuring "everyone ever-lastingly tested." His was an enormous contribution to the belief that educational affairs can be conducted more effectively if we measure how each student differs and does not differ from the others.

Psychology had been recognized only a few years before as a separate discipline, somewhere between philosophy and physiology. Its mission was to explore the connection between the human mind and the human body. With his *On the Origin of Species*, Galton's cousin Charles Darwin[8] helped stimulate these new psychologists to ponder the relative importance of

[8] Darwin, C. R., 1859. On the origin of the Species. First presented at a lecture by colleagues.

heredity and environment as causes of behavior. Galton was very much a hereditarian:[9]

> *I have little patience with the hypothesis occasionally expressed, and often implied, especially in tales written to teach children to be good, that babies are born pretty much alike, and that the sole agencies in creating differences between boy and boy, and man and man, are steady application and moral effort. It is in the most unqualified manner that I object to pretensions of natural equality. The experiences of the nursery, the school, the University, and of professional careers, are a chain of proofs to the contrary. I acknowledge freely the great power of education and social influences in developing the active powers of the mind, just as I acknowledge the effect of use in developing the muscles of a blacksmith's arm, and no further. Let the blacksmith labour as he will, he will find there are certain feats beyond his power that are well within the strength of a man of herculean make, even although the latter may have led a sedentary life. Some years ago, the Highlanders held a grand gathering in Holland Park, where they challenged the English to compete with them in their games of strength. The challenge was accepted, and the well-trained men of the hills were beaten in the foot-race by a youth who was stated to be a pure Cockney, the clerk of a London banker.*
>
> *Everybody who has trained himself to physical exercises discovers the extent of his muscular powers to a nicety. When he begins to walk, to row, to use the dumb bells, or to run, he finds to his great delight that his thews strengthen, and his endurance of fatigue increases day after day. So long as he is a novice, he perhaps flatters himself there is hardly an assignable limit to the education of his muscles; but the daily gain is soon discovered to diminish, and at last it vanishes altogether. His*

[9] Galton, F. 1883. *Inquiries into Human Faculty.* Sagwan Press.

maximum performance becomes a rigidly determinate quantity. He learns to an inch, how high or how far he can jump, when he has attained the highest state of training. He learns to half a pound, the force he can exert on the dynamometer, by compressing it. He can strike a blow against the machine used to measure impact, and drive its index to a certain graduation, but no further. So it is in running, in rowing, in walking, and in every other form of physical exertion. There is a definite limit to the muscular powers of every man, which he cannot by any education or exertion overpass.

This is precisely analogous to the experience that every student has had of the working of his mental powers. The eager boy, when he first goes to school and confronts intellectual difficulties, is astonished at his progress. He glories in his newly-developed mental grip and growing capacity for application, and, it may be, fondly believes it to be within his reach to become one of the heroes who have left their mark upon the history of the world. The years go by; he competes in the examinations of school and college, over and over again with his fellows, and soon finds his place among them. He knows he can beat such and such of his competitors; that there are some with whom he runs on equal terms, and others whose intellectual feats he cannot even approach. Probably his vanity still continues to tempt him, by whispering in a new strain.

It tells him that classics, mathematics, and other subjects taught in universities are mere scholastic specialties, and no test of the more valuable intellectual powers. It reminds him of numerous instances of persons who had been unsuccessful in the competitions of youth, but who had shown powers in after-life that made them the foremost men of their age. Accordingly, with newly furbished hopes, and with all the ambition of twenty-two years of age, he leaves his University and enters a larger field of competition. The same kind of experience awaits him here that he has already gone through. Opportunities occur --

> *they occur to every man -- and he finds himself incapable of grasping them. He tries, and is tried in many things. In a few years more, unless he is incurably blinded by self-conceit, he learns precisely of what performances he is capable, and what other enterprises lie beyond his compass. When he reaches mature life, he is confident only within certain limits, and knows, or ought to know, himself just as he is probably judged by the world, with all his unmistakable weakness and all his undeniable strength. He is no longer tormented into hopeless efforts by the fallacious promptings of overweening vanity, but he limits his undertakings to matters below the level of his reach, and finds true moral repose in an honest conviction that he is engaged in as much good work as his nature has rendered him capable of performing.*

Galton, the hereditarian, said that each student discovers his station among his fellows. An environmentalist would not disagree. Galton was describing today's castes as well as yesterday's. No educational system, old or new, progressive or classical, is characterized by mobility up and down the ranks-in-class.

In America, across those years, the public schools were celebrated as the agent of self-improvement. The industrious student, with the guidance of a sensitive teacher, could lift himself to higher station. A teacher's sensitivity was defined partly as knowledge of the child's mental capacities. The mental test movement grew, from an effort in 1900 to discriminate between those children who could and could not profit from formal schooling, to an effort in 1930 to identify the scholastic promise of every child.

Galton's anthropometric techniques did not prevail. As developed by James McKeen Cattell, they measured intelligence by measuring such things as speed of arm movement, reaction time to sound stimuli, and memory for random letters heard once. It did not work. Rather, the techniques of Alfred Binet, measuring "higher mental processes" based on experiences common to most children, did prove to be a reliable base for predicting later achievement. Two generations later, each child owned a momentary mental age and a durable intelligence quotient, indicators of his standing in class and among all children everywhere.

Galton had pointed out in 1869 the regularity of the distribution of human differences, and that these very differences could be used as a dependable scale of measurement:

> *The range of mental power between -- I will not say the highest Caucasian and the lowest savage -- but between the greatest and least of English intellects, is enormous. There is a continuity of natural ability reaching from one knows not what height, and descending to one can hardly say what depth. I propose in this chapter to range men according to their natural abilities, putting them into classes separated by equal degrees of merit, and to show the relative number of individuals included in the several classes. Perhaps some persons might be inclined to make an offhand guess that the number of men included in the several classes would be pretty equal. If he thinks so, I can assure him he is most egregiously mistaken.*
>
> *The method I shall employ for discovering all this, is an application of the very curious theoretical law of "deviation from an average." First, I will explain the law, and then I will show that the production of natural intellectual gifts comes justly within its scope.*
>
> *The law is an exceedingly general one. Adolphe Quetelet, the Astronomer-Royal of Belgium, and the greatest authority on vital and social statistics, has largely used it in his inquiries. He has also constructed numerical tables, by which the*

necessary calculations can be easily made, whenever it is desired to have recourse to the law.

... Suppose a million ... men to stand in turns, with their backs against a vertical board of sufficient height, and their heights to be dotted off upon it. The board would then present the appearance shown in the diagram (not included here). The line of average height is that which divides the dots into two equal parts, and stands, in the case we have assumed, at the height of sixty-six inches. The dots will be found to be ranged so symmetrically on either side of the line of average, that the lower half of the diagram will be almost a precise reflection of the upper. Next, let a hundred dots be counted from above downwards, and let a line be drawn below them. According to the conditions, this line will stand at the height of seventy-eight inches. Using the data afforded by these two lines, it is possible, by the help of the law of deviation from an average, to reproduce, with extraordinary closeness, the entire system of dots on the board.

The number of grades into which we may divide ability is purely a matter of option. We may consult our convenience by sorting Englishmen into a few large classes, or into many small ones. I win select a system of classification that shall be easily comparable with the numbers of eminent men, as determined in the previous chapter. We have seen that 250 men per million become eminent; accordingly, I have so contrived the classes in the following table that the two highest, F and G, together with X (which includes all cases beyond G, and which are unclassed), shall amount to about that number -- namely, to 248 per million.

[Here should follow a highly populated table that I am leaving out, entitled: *Classification of Men According to their Natural Gifts.* Omitted are the estimated numbers of men of each age group and each grade of natural ability then living in England and Wales.]

Example: The Class F contains 1 in every 4,300 men. In other words, there are 233 of that class in each million of men. The same is true of Class f.

It is an absolute fact that if we pick out of each million the one man who is naturally the ablest, and also the one man who is the most stupid, and divide the remaining 999,998 men into fourteen classes, the average ability in each being separated from that of its neighbours by equal grades, *then the numbers in each of those classes will, on the average of many millions, be as is stated in the table. The table may be applied to special, just as truly as to, general ability. It would be true for every examination that brought out natural gifts, whether held in painting, in music, or in statesmanship. The proportions between the different classes would be identical in all these cases, although the classes would be made up of different individuals, according as the examination differed in its purport.*

A Momentous Idea

The idea of precisely describing an individual's standing in a particular population of individuals, and of using the variation in individuals as a unit of measurement, was as great an invention as that of gunpowder. And perhaps as benevolent.

It is not evident that education is improved when there is knowledge of a student's standing in his group. The fact that a scientifically respectable technology of testing for individual differences is available does not make it benevolent to the learner, or even useful to the teacher. Galton and his followers looked more for the discriminations that are possible than for those that are useful -- with some justification. We often do not know what will be useful before we have it in use. They built millions of tests, most without external validity.

Tests have purposes other than to indicate students' relative standing. They do confirm, and sometimes nicely contradict, a teacher's expectation of what a child can do. It is apparent that testing activities are used (Galton's indignation notwithstanding) to spur youngsters to greater effort. Scores continue to be the basis for granting educational opportunity to some and denying it to others.[10]

But the purpose here is not to evaluate all of testing. It is to question the use -- by educators and particularly by educational measurements people -- of the concepts derived from the individual difference scale, namely, class-ranks, grade-point-averages, percentile ranks, and correlational statistics.

Many -- Bob Glaser and Anthony Nitka, (See Shub, 1977)[11] for example--have criticized the prevalent "norm-referenced" measurement on the ground that there the subject matter of particular test items is only important as representing more general knowledge or skill, not important in itself. They prefer test items to be specifically related to instructional objectives and interpreted individually more than collectively. Many believe the important information to be gained is not how someone stands with reference to a group but whether or not this someone has mastered the specified learning task. Two

[10] Thorndike, R. L., 1951. *Educational Measurement*. American Council on Education

[11] Shub, A. N., 1977. A Descriptive Exploratory Analysis of Some Issues in Criterion-Referenced Measurement with Possible Application to a Diagnostic-Prescriptive System for Developing Measurement Competency for Prospective Teachers. (doctoral dissertation) Chicago: Loyola University.

issues are thus confounded: (1) whether to orient the curriculum to the general content-coverage deemed suitable for unspecified transfer of training or to orient it to the mastery of specified knowledge and skill; and (2) whether to interpret test scores in terms of the specific content of each test item or on the basis of inter-personal correlation of the total test performance with other educational performance, i.e., to interpret scores in terms of individual differences. The two issues are related but deserve separate attention. Hewing to the Galton side, let us concentrate on the relevance of orderings of students for the conduct of instruction.

With years of practical experience and correlational research behind us we know much about the durability of student orderings: e.g., we know those students standing high in the group in mathematics performance today are likely to be those standing high in the group in language performance tomorrow, and so on. The basis for good prediction is here in our grasp; but seldom is it important for instructors to predict later standing.

The *basic claim* of individual-psychologists in education is that knowledge of a student's standing is a basis for differential treatment that will increase learning. Students scoring above a certain point might be assigned one teaching mode, those below, another. Or that for every increment of superior test performance, a variation in teaching. Or each child tutored on the basis of his unique complex of measured attributes. With a promise of increased learning.

One would expect, for example, if the claim were true that tests of "reading readiness," administered in many schools prior to instruction in reading, would be validated on the basis of the different activities that might be assigned to the children who perform differently on the tests. But they are validated only on whether or not the children at a subsequent time are rank ordered in reading skill similarly to the test's rank ordering.

Reading readiness tests tell us something about what are reasonable and unreasonable performance expectations for these children, but not whether it is best to commence or postpone reading instruction.

The claim that it helps to know individual differences is implied over and over in the first two chapters of *Educational Measurement*.[12] One chapter is "The Functions of Measurement in the Facilitation of Learning" by Walter Cook and the other is "The Functions of Measurement in Improving Instruction" by Ralph Tyler. Both authors noted the countless ways in which individuals differ and the fai1ings of common sense to appreciate those differences. But neither demonstrated the validity of the claim that better knowledge of individual differences will in fact lead to better teaching and learning.

Glenn Bracht (studying aptitude-treatment interactions[13]) and Urban Dahllöf (studying ability grouping[14]) searched the literature and gathered new data, but found no support for the claim. Lee Cronbach and Dick Snow (1969) looked widely too and concluded that *"... there are no solidly established aptitude treatment interactions even on a laboratory scale and no real sign of any hypothesis ready for application and development."*[15]

Phil Jackson[16] put the question simply, "*Is there a best way of teaching Harold Batemen*?" and answered it "No". He acknowledged that what a teacher does for Harold at any time could often be recognized as one of the better or poorer things to do, but reasoned that information such as individual

[12] Thorndike, R. L., 1951. *Educational Measurement*. American Council on Education.

[13] Bracht, G., 1970. Personal communication.

[14] Dahllöf, U., 1971. Personal communication.

[15] Cronbach, L. J., & Snow, R. E., 1977. *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

[16] Jackson, P., 1986. *The practice of teaching*. New York: Teachers College Press.

standings in test performance will probably never be an important or even statistically valid base for teachers making the choice.

We do not find support for the individual differences claim by examining the experience of classroom teachers. Teachers frequently teach different students differently and they are very much aware of different abilities and interests. Few rely (or advise others to rely) on formal knowledge of relative standing. The lore of teaching reveals little of how to teach the uppers and the lowers differently except for how far back to start and how fast to go.

The experienced teacher does adjust the coverage and pacing of instruction to assist individual learners. Recitations, quizzes and standardized tests do help a teacher stay aware of changes needed in instruction. The most effective teachers do measure, in many ways. But this all does not support the common claim that additional knowledge or use of individual differences will improve classroom instruction and increase the opportunity for learning.

We still do not know whether or not youngsters will benefit more by admission or non-admission, or more by one form of instruction than another. If the truly individualistic point of view is taken, that is, that schools should offer the best possible educational opportunity to each individual student rather than that the schools should help society allocate the limited supply of educational resources, opportunity, and privilege to the student body, then at this time there is very little direct or indirect educational value in the knowledge of how a child compares to other children in aptitude or achievement.

The point of view challenged here is the one that sustains the use of percentile ranks, rank in class, or ability grouping, and also the point of view that sustains much of psychometric theory and test development. Reliability and validity of educational measures are usually defined in terms of the correlation between the orderings of students. Instructional research findings, even in the most fastidiously designed studies, are often interpreted in terms of variance explained, the differences in *individual standing* that can be accounted for in terms of the treatments under scrutiny. Educational treatments are important because they change opportunities to learn and because they improve or hurt the performance of the individual learner -- not because they correlate with "standing in the group. "

Validity is crucial. And "validity for what?" needs to be asked again and again. Many techniques are validated for discriminating among individuals, but for nothing else. When we are trying to discover how well a child can read, or when we observe a child trying out a new workbook, or when we scrutinize the match between stated objectives and actively pursued objectives, our observations need to be valid, but the concept of individual differences is irrelevant. Valid measuring is getting information that would be confirmed under alternate methods of inquiry. In education most important measures are valid if they yield information that competent educators and researchers could observe for themselves, were conditions to permit it. Changes in phenomena over time and differences in personal perspective make validity estimates difficult, but the solution does not involve an appeal to individual-difference measurement. Only occasionally will it be appropriate to use stability of interpersonal ordering to validate measurements. And that of course will be on those occasions we really want to know how an individual person compares with others.

Researchers should be more concerned with what educators and laymen want to know. Sometimes they do want to know about ranks and about the many statistics derived from the distribution of talent in the population. But often the attributes of others in the population are quite irrelevant.

Francis Galton did not create the anthropometric station because he wanted to keep things stationary. He wanted to know how people are different, and why. He wanted to change things. He created the term *eugenics.* He wanted to extend the privileges *and* responsibilities of society to those “superior” in the talents that his kind of people admired.

Historian Clarence Karier[17] claimed that the primary effect of educational testing is to allocate privilege in our society to "merit" rather than on any egalitarian basis, legitimatizing favoritism to those-most-dear-to-those-in-charge, and, keeping others in their place. (I wish I had listened sooner. Bob).

[17] Karier, C. J., 1986. *The Individual, Society, and Education: A History of American Educational Ideas, Second Edition*. Urbana: University of Illinois Press

1973

Wrote this while still in Sweden and just back from Copenhagen. Now, rereading, we are a couple of generations removed. But even as we have become more wary, we remain fascinated by measurement. Like some of the carpenter's tools, some of ours have sharp edges. Two of the carpenter's tools were the hawk and the handsaw.

A Hawk and A Handsaw

Education, like Denmark and other sovereignties, is made up of many principalities. We even call some of our noblemen and noblewomen "principals." A few of the others we call "specialists in measurement."

The work of the specialist in measurement is to observe and record teaching and learning. Observations are his products. Of course, the measurements specialist is not the only one who observes educational events and artifacts. Everyone does. Teachers and students and commissioners do an enormous amount of observing and recording -- but they each have other duties. It is the specialist in measurement whose primary duty is to discern the educational events happening.

It may be a mistake to have such division of labor. The specialists, here as elsewhere, get touchy about jurisdictions (the realm of the duchy) and about the pomp and circumstance of authentic observations (you remember what they said about Professor Harold Hill: *... "but he doesn't know the territory!").* The counselors and custodians are too often told to mind their own business and worse, too often tell others they have their own business to tend to -- when the business of teaching and learning and managing the schools is everyone's business. Every professional educator, at one time or another, sometimes every minute of most days, is the principal measurement person on duty. The work of measurement is their work too. It is everyone's job to see what exists as education.

If anything exists it exists because someone in some way has sensed it, thought of it, grabbed it, been grabbed by it, dreamed of it.

To most people the important things to measure are not those that are experienced by only one person, but are public, shared and sharable by many. Of course, many of the things most precious are experienced in a non-public, unique way by each person. And I see no reason to exclude from *existence* those things that are apparent only to one person. It may be important to try to find out "who all" saw it and "who all did not."

Edward Thorndike, sometimes said to be the father of educational psychology, urged us to attend more to the quantities: "Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality." [1] But of course one cannot know something at all if he knows only a few well-quantified attributes: "She is a girl, age 9, in the fourth grade, scholastic aptitude percentile rank 66 on national norms, spent 38 minutes on Lesson D7.32, scored 82% correct on Criterion Test D7.32, checked Green-Box-2 on Preference-to-Proceed." [2]

A more elementary description should take priority over quantitative refinement. The prior question is: What exists at all? Something cannot exist unless someone realizes its presence. The first basic measurement is the realization of existence, the difference between some and none. When someone experiences something, he says, in one kind of talk or another: "It is not zero." He has made a measurement. When we say,

[1] Thorndike, E. L., 1904. *An Introduction to the Theory of Mental and Social Measurements.* Forgotten Books.

[2] Stake, R. E., 1973. Made up for this essay.

"That is a brick*."* a measurement has been made. When she picks the last petals of a daisy, saying: "He loves me, he loves me not," a measurement has been made. (We will talk about validity later.) When we say: "There is a spirit of free expression in this room," a measurement has been made. Basic measuring is the discerning of existence: one, not zero,

There is a common belief that measurement is a form of mathematics. Its development and analysis of observations often require mathematical techniques, but measurement appears in simple forms that most people do not think of as mathematical. Educational measurement may have gotten more mathematical than it need be, perhaps because many of us measurements people are former mathematics teachers. We may have been carried away by Thorndike's creed,[3] as quoted here:

> *Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality. Education is concerned with changes in human beings; a change is a difference between two conditions; each of these conditions is known to us only be the products produced by it -- things made, words spoken, acts performed, and the like. To measure any of these products means to define its amount in some way so that competent persons will know how large it is, better than they would without measurement. To measure a product well means so to define its amount that competent persons will know how large it is, with some precision, and that this knowledge may be conveniently recorded and used. This is the general Credo of those who, in the last decade, have been busy trying to extend and improve measurements of educational products.*

[3] Thorndike, E. L., 1904. *An Introduction to the Theory of Mental and Social Measurements*. Forgotten Books.

We have faith that whatever people now measure crudely by mere descriptive words, helped out by the comparative and superlative forms, can be measured more precisely and conveniently if ingenuity and labor are set at the task. We have faith also that the objective products produced, rather than the inner condition of the person whence they spring, are the proper point of attack for the measurer, at least in our day and generation.

This is obviously the same general creed as that of the physicist or chemist or physiologist engaged in quantitative thinking -- the same, indeed, as that of modern science in general. And, in general, the nature of educational measurements is the same as that of all scientific measurements. The nature, purposes, and general methods of educational products.

Well, in 1973, it is a new day and a new generation manifestly so in that at this writing Edward Thorndyke's son Bob is a measurements man and president-elect of the American Educational Research Association. In this new day there is increased skepticism that greater precision leads to more valid and useful results. The creed is challenged.

In this day, many educators have responsibilities stretched to distant and unfamiliar places. They are strangers to the territory. They are dependent on formal communication media and measurement information to see the word-products and act-products that they cannot see for themselves. So today the primary measurement obligation is not to increase the precision of the statement of how large something is but to describe it, crudely perhaps, helped out by the comparative and superlative forms.

What is the comparative form? Hamlet said: “I am but mad north northwest; when the wind is southerly, I know a hawk from a handsaw.” To most Danes, to most audiences, a hawk and a handsaw are little alike. To some carpenters they both are tools. Hamlet tied the comparison to sanity. Our basic thought processes require such comparisons. Our basic measurement processes emphasize discrimination among different things.

When two things are compared, we observers see a difference or no difference, a big difference or little difference. We may see some attributes that are similar and some attributes that are really different. Noting the hue of neckties or the cry of babies we may report: "This one is louder." To report the result of a comparison calls for a description of differences. The quantification in this measurement is the recognition and reporting of inequalities. The second act of measuring is the discerning of differences.

What is the superlative form? William Shakespeare had Rosencrantz describe King Claudius: *"It is a massy wheel fixed on the summit of the highest mount, to whose great spokes, ten thousand lesser things are mortis'd and adjoin'd; which when it falls, each small annexment, petty consequence, attends the boisterous ruin. Never alone did the King sigh, but with a general groan."*

When many lives are together compared, and one stands out, it is the superlative. One day is longest, one child most troublesome, one state plan most comprehensive. Or there are *the few* that are longest, most troublesome, or most comprehensive. Many others are observed to be ordinary; no more than a few are superior or inferior. The reference group may be carefully specified or only vaguely implied. The third basic measurement is distinguishing the outstanding few from the many.

Measurement technology is needed in this generation too. There still must be distinctions between good and bad observing and between good and bad reporting. And often we will rely on the inner condition of persons to judge what is good and bad. With the wind in the south Hamlet said: "…for there is nothing either good or bad, but thinking makes it so ...” We continue to need dependable techniques for getting good judgments about observations and reports.

Some measurements specialists are reluctant to accept subjective opinions as the criteria for successful measurement. They would have us work with more exact scales. But it is our lot in education to work where discriminations can be both gross and important, such as saying: “The bureaucracy of this curriculum-development project is growing.” and “This child cannot comprehend what he is reading any better than he could a year ago, even though he now is reading more difficult materials.” We make these statements partly in terms of whether our co-workers and clients consider perceived differences to be non-negligible (a subjective opinion) and relevant (another subjective opinion).

Many educators are reluctant to admit that they have responsibilities for measurement. They have been made to feel that anyone who does must be competent in mathematics and the technology of standardized testing. Few feel that *they* are ready to apply mathematical terminology and analysis to educational problems. And, moreover, many are dismayed by statistical portrayals of education and prefer not to be associated with them. So they, unfortunately, are reluctant to acknowledge their measurement responsibility.

We who teach testing, measurement, and statistics courses in Colleges of Education and elsewhere are particularly to blame. We teach highly quantitative courses, though we need not. We teach what we enjoy teaching. Some of us enjoy teaching courses that remove or discourage students with *low quantitative aptitudes* from the pursuit of higher degrees in education. We try to teach a course that our graduate faculty colleagues will consider rigorous -- whether or not the processes and standards thus emphasized are the basic ones for educational measurement.

These courses should be taught. Specialists in measurement should be required to take them. But those not specializing in measurement should not be required to do so and they should not be led to believe that by taking them they are learning basic measurement as applied to education. They and their advisors and a legion of curriculum committees should realize that these courses do not go far toward developing the technical skills needed for professionally observing and analyzing educational phenomena.

Observation, perceptual discrimination, comparison, and referencing are the basic measurement processes. The data are observations of existence, equality and inequality, ordinariness and outstandingness.

Their products are comprehensive descriptions. No educator should be persuaded that it is highly important to study the procedures for developing tests and analyzing scores if he has not yet learned the procedures for making, corroborating, analyzing and synthesizing observations. This is true for the person studying to be a better practitioner and it is true for the person studying to be a better researcher. Education and educational research could do without testing, but they cannot do without observation.

Edward Thorndike spoke of knowing something thoroughly. So little of education do we know thoroughly. With much of it we have so meager an acquaintance. We seek differentiation. How will measurement help us? ... that is the question.

1973

I wrote this during my sabbatical stay at the University of Göteborg, representing the provocation of thought from talks with Urban Dahloff, Ulf Lundgren, Kjell Härnqvist, Ference Marton, and Erik Wallin. I had gone there to study how research might be more useful in a society more rational than the American. I had been eighteen years in the thrall of quantitative research methods, groping for ways to make my inquiry and teaching better. Sweden drew me to a new literature, making me ready for immersion in qualitative inquiry, particularly case studies in East Anglia and Illinois. Back home I had long discussions with Stephen Kemmis and other colleagues, voices from near and far.

Voices

And then there is the story of the empiricist who walked among people to show his luxuriant clothes, and all were quiet so as not to reveal their inability to appreciate such trappings, until a child cried. out, "... but he doesn't have anything on!"

Now I personally have not concluded that as a group, empiricists, those folks who rely on observations and measurement as a basis for decisions, are less decently clothed than the rationalists, those folks who rely on thinking it logically. But as a rationalist I wonder about the claims we empiricists make as to how well our measurements serve others.

I recall James Thurber's[1] story of Princess Leonor. She lay ill longing for the moon. The King's wizards could only recite a list of their past findings and declare the impossibility of the Princess' request. But the Court Jester talked to Leonor, made

[1] Thurber, J., 1943. Many Moons. Harcourt Brace.

suggestions, and found in a gold medallion, the moon to restore her health.

The relief was hidden there in Leonor's mind. Those who would help needed to speak to her, and to know the other voices that spoke to her. Alas, it offends the social scientist to be told that better answers to social problems may lie hidden in the minds of the afflicted than can be brought to light by research. Offended or not, the researcher will sometimes gain better results by helping practitioners find their own solutions rather than finding solutions for them.

What do we know about making sick education well? In my discussions with reviewers, Gene Glass,[2] Barak Rosenshine[3] Daniel Kallos[4] and Ed Short,[5] we don't know much. Nate Gage[6] concluded that "... *positive results remain hard to come by.*" They were talking about the formal knowledge we have accumulated to improve ineffective teaching and learning. More broadly, the National Science Foundation reported in 1969[7]: "*Too few mechanisms for translation of social scientific understanding into societal benefit have been institutionalized so as to assure this process.*"

One of the more unsupportive voices is that of the highly respected philosopher of science Thomas Kuhn, who told

[2] Glass, G. V., 1971. Educational Knowledge Use. *The Educational Forum. 36*, 21-29.

[3] Rosenshine, B., 1997. Advances in research on instruction. In J.W. Lloyd, E.J. Kameanui, and D. Chard, Editors, *Issues in educating students with disabilities.* Mahwah, N.J.: Lawrence Erlbaum: pp. 197-221.

[4] Kallos, D., 1973. On educational scientific research. Unpublished paper. Lund University: Sweden: Pedagogiska Institutionen.

[5] Short, E. C., 1970. A review of studies on the general problem of knowledge production. College of Education, University of Toledo, (ERIC: ED 055 023).

[6] Gage, N., Personal discussions.

[7] National Science Foundation, 1969. The Annual Report of the National Science Foundation, 1969.

American Educational Research Association executive director, Richard Dershimer, in 1970[8]:

> *I'm not sure that there can now be such a thing as really productive educational research. It is not clear that one yet has the conceptual research categories, research tools, and properly selected problems that will lead to increased understanding of the educational process. There is a general assumption that if you've got a big problem, the way to solve it is by the application of science. All you have to do is call on the right people and put enough money in and in a matter of a few years you will have it. But it doesn't work that way, and it never will.*

Kuhn implied that to be productive, research must produce formal understandings. He did not see that happening in Education. If all these voices are correct, should anybody be doing educational research? That is not a realistic question. Even if the costs are high and even if the benefits were negligible, there would be research. First, people will follow their curiosities. Second, the preparation to do research is too large an investment to abandon. And third, people demand that claims and actions be justified. Even by their very existence research reports are "justification" in a technological society. Studies and documents are "life blood" in bureaucracies and corporations. There will be educational research. The question is *for what and how*, not of *whether*.

One of the less pessimistic voices is that of the highly respected developer of individualized learning materials, Bob Glaser, who said in his 1973 presidential address to American Psychological Association educational psychologists: [9]

[8] Dershimer, R.,1969. Personal communication.

[9] Glaser, R., 1973. The new aptitudes and adaptive education. Vice Presidential Address. Annual Meeting of the American Psychological Association.

> *The behavioral and social sciences are at a point in their development where they absolutely require the direction and disciplining effects that come from contact with real-world problems. Fortunately, this is more possible than ever in the light of the growing openness of society toward innovation and experimentation. What knowledge and theory have been accumulated now need the elaboration and correction that can result from such engagement. The sequence from basic research, to applied research, to development, to practice and application on which most of us were weaned is no longer applicable if, in fact, it ever was. ... [We need a three-way] interactive mode of operation between application, technology and basic science.*

Many people see, as Glaser does, the payoff of tomorrow's research dependent upon its communication with practitioners and technologists.

A large social investment in research (in careers, institutional attention, and money) can be justified if it yields scientifically-robust understandings -- but in other ways as well. Research provides data and theories, but also concepts and metaphors, for the discussion of educational problems -- and may stimulate that discussion. Moreover, modestly-budgeted research is a worthwhile investment if it is an enactment of our best hunches on improving practice, a justifiable sustenance to those striving to provide a better social service. But the common expectation is that research will justify itself, or it will not, in terms of the knowledge it produces and the practices it improves.

Perhaps because education is so "knowledge-oriented" many of us have an enduring expectation that *knowledge* can make sick education well. In a review of the impact of curriculum research Ed Short wrote that

> *a number of researchers have redefined the scope of the phenomena and have conceived it, not as a problem of "research into practice" but as one of "knowledge production and utilization."*

The idea that the two enterprises, *knowledge production* and *knowledge utilization* rise high and reach out to each other, perhaps to form a golden arch, is an attractive picture, particularly to someone drawing up an accountability system or reform plan. But according to Short's findings, and as many see it, the two are seldom reaching for each other, much less embraced. Teachers and administrators have little use for the information gathered. Researchers have little interest in the web of personalistic and crisis-like problems of the day.

And Short's reformulation statement may be as disappointing as the arch itself. The statement diminishes the concept of Practice by its attention to Knowledge. The possibility that practice could not be aided by formal knowledge is obscured. Some of us need to continue to ask, *"Research into Practice, how?"*

We live in a society with much expected of inquiry. Better knowledge, better products, better work methods, less work, a better life. And we are rewarded, with heads full of information, tool kits full of hardware, closets full of trappings, and ever-expanded curiosities. Inquiry succeeds, not always by aim -- but certainly not indiscriminately. Knowledge research is not product research. Theory testing is not problem-solving. From any research, we might gain a new theory, a new set of facts, an explanation, an understanding, a proof, an interpretation, a resolution of dilemma, a justification, an inspiration, ... and some will come when we least expect them, but each is more likely to come from particular ways of inquiry, and less likely to come from others. Successful researchers tailor their inquiry to their purposes.

Psychologist Lee Cronbach and Philosopher Pat Suppes[10] distinguished between *conclusion-oriented research*, that aimed to further our general understanding, and *decision-oriented research*, that aimed to get a particular job done or problem solved. Gene Glass and Blaine Worthen[11] [12] identified nine characteristic ways in which evaluative *inquiry* differs from *evaluative* inquiry.

The most powerful distinction in research purpose that I see is very close to those, but I will continue to speak of a choice between practice and knowledge,

research → practice or *research → knowledge*

I chose not to use the distinction between conclusions and decisions partly because it confuses me and partly because I want to acknowledge the often insignificant role that abstract knowledge and deliberated decisions play even in very successful professional practice.[13] And I think that it is useful to see evaluative research as something special, but I want to emphasize all research contributions to practice, evaluative and otherwise.

The end-product here is the well-being of practice. Perhaps that means changed practice, perhaps unchanged practice, but always practice. The end-product is not knowledge

[10] Cronbach L. J. and Suppes, P., 1969, *Research for tomorrow's schools. Disciplined inquiry for education.* Toronto: Macmillan.

[11] Glass, G. V. and Worthen B. R., Personal communication.

[12] Glass, G. V., 1972. The wisdom of scientific inquiry on Education. *Journal of Research in Science Teaching, 9*, 1, 3-18.

[13] It is a common belief among researchers that school practice would be improved if it were more rational and deliberative. Such a belief itself of course does not justify a research design whose findings are useful only if practitioners act as rationalistic decision-makers, or in any new and unfamiliar way. A researcher's advocacy of rationalism needs to be recognized as separable from his inquiry services, and sometimes to be arrested.

about practice. Nor is knowledge necessarily an *intermediate* product. It is not assumed here that in order to improve practice the practitioners must increase their knowledge or awareness of practice. Knowledge may help, awareness may help, but in each case that remains to be seen and to be empirically verified.[14]

A major importance of the distinction between

research → practice and *research → knowledge*

is that different criteria will be used to judge the worth of the two researches. Practice-bent research is intrinsically good if it stimulates well, and intrinsically good if it results in practice of a high quality. Knowledge-bent research is intrinsically good if it asks good questions well, and extrinsically good if it results in knowledge of a high quality. Of course, the standards need operationalization, but the point here is apparent: that the two kinds of research require criteria of success distinct and appropriate to their separate purposes.

This distinction is considered trivial by those who believe that practice or corrective action is essentially knowledge-based, and that the essential formulation for applied research and reform is

research → knowledge → practice

Surely this *is* sometimes the right model for guiding research. And surely, in retrospect, changes in practice can almost always be attributed to changes in knowledge. But it will be more useful sometimes for research to be designed so that

[14] Research toward such an end is sometimes called *action research*. I was tempted to refer to this paper as my theory of action, but I decided that that implies that practice needs to be changed. A researcher desiring to be of help should hesitate to say that change is needed more than protection.

research → communication → practice

If there is communication, some knowledge will be communicated, but the knowledge itself may be of little moment in the dynamics of reconsidering practice. Communication -- the dialogue, the acts of discussing the problems, drawing upon other experience, subjecting opinions to scrutiny, talking of alternative expectations and criteria -- can be the important precursor to action. Research then should be examined in terms of its contribution to communication independent of its contribution to knowledge.

It is interesting to look backward step by step at the etiology of practice. We should keep in mind that we are primarily interested in the activities of someone responsible for part or all of an educational system -- a classroom, a district, a college, or a national agency. This person is moved to act, to continue an activity, to refrain from acting, or to resist action only when he or she has the power to do such a thing, and when sufficient external demand or internal conviction arises.

demand + conviction + leverage → action (practice)

Research can contribute to practice by changing the leverage or conviction, or the demand from students, authorities, citizens and others. Product research, curricular innovations, and administrative operations-research seek to increase leverage, to increase the number of options, and to find new ways of allocating resources. Little educational research is designed to change the external demand on the decision maker, though consumer research and evaluative studies could be. Much of educational research is intended to influence the conviction of the practitioner.

It is implicit in most of the designs that the decision maker is a free agent, capable of generalizing from remote happenings and of applying or adapting findings to local circumstances. There is an expectation of appetite for explanations of the dynamics of the system and relationships that account for the uniqueness of individual persons and events. These abilities and appetites do influence a practitioner's conviction but they are secondary to personal effects, such as needs for security, autonomy, and prestige, and to such daily obligations as keeping order in the hallway and resolving conflicts between department heads. The practitioner is only a little bit free to embrace new understandings.

What is missing in those designs is the realization that action is precipitated by voices, the voice of experience, of reason, of camaraderie, of adulation, of indignation, of aspiration. These voices are seldom still.

And still it may be said that conviction to act or to forbear action is a product of understanding and motive.

understanding + motive → conviction

Persons make a different commitment to something when they change either their understanding of it or the value they hold for it. Perhaps one cannot change without the other. In the matter of class size, for example, a teacher may be convinced that the smaller the class the greater the learning opportunity. To be of service to this teacher the researcher may work on the conviction (perhaps eventually to reinforce it, perhaps to weaken it) by studying and talking about the effects on learning of changing class-size. A change in conviction may occur because the teacher becomes more understanding about the teaching-learning processes in groups of different size and/or because the "costs" and "benefits" of altering class size require changed measurements. The teacher might become persuaded that ability grouping is almost unavoidable when big classes are broken into small, and that the quality of education (both in content and in experiential value) is almost unavoidably reduced for the slower learner separated from faster learners. With this small shift in motive and understanding may rise an increased conviction to accept the more impersonal less easily managed, large classes.

Philosophers of science have disagreed as to the similarity of the understandings that spin out from C. P. Snow's[15] two cultures, the natural sciences and the humanities. Natural scientists have sought to explain the happenings of the natural world, independent of supernatural or human purpose. Their ally, the positivistic philosopher of science claims that such a style of research is an appropriate model for all inquiry. Humanistic scholars have sought a comprehension, even an apprehension, of human experience, with attention to purpose, empathy and dialogue. Their ally, the anti-positivistic philosopher of science, sometimes called an idealist, sometimes a proponent of "hermeneutics," contends that *understanding* is different from *explanation*.

A careful observer of this scene, philosopher George Hendrik von Wright, described the humanistic researchers not as soft-headed opponents of mechanization and rigor, but as fully legitimate inquirers with a most respectable lineage. In *Explanation and Understanding*, [16] he described their spokesmen in these extensive words

All these thinkers[17] *reject the methodological monism of positivism and refuse to view the pattern set by the exact natural sciences as the sole and supreme ideal for a rational*

[15] Snow C. P., 1959. *The Two Cultures.* The Rede Lecture.

[16] von Wright, G. F., 1971. *Explanation and Understanding*, Cornell University Press, pages 5-7.

[17] Thinkers like Droysen, Dilthey, Simmel, and Max Weber, plus Windelband and Tickert of the neo-Kantian Baden School.

understanding of reality. Many of them emphasize a contrast between those sciences which, like physics or chemistry or physiology, aim at generalizations about reproducible and predictable phenomena, and those which, like history, want to grasp the individual and unique features of their objects. Windelband coined the label nomothetic for sciences which search for laws, and "ideographic" for the descriptive study of individuality.

The anti-positivists also attacked the positivist view of explanation. The German historian-philosopher Droysen appears to have been the first to introduce a methodological dichotomy which has had great influence. He coined for it the names explanation and understanding, in German "Erklaren" and "Verstehen." The aim of the natural sciences, he said, is to explain; the aim of history is to understand the phenomena which fall within its domain. These methodological ideas were then worked out to systematic fullness by Wilhelm Dilthey. For the entire domain of the understanding method he used the name Geisteswissenschaften. There is no good equivalent in English but it should be mentioned that the word was originally coined for the purpose of translating into German the English term, moral science.

Ordinary usage does not take a sharp distinction between the words explain and understand. Practically every explanation, be it causal or teleological or of some other kind, can be said to further our understanding of things. But understanding also has a psychological ring which explanation has not. This psychological feature was emphasized by several of the nineteenth-century anti-positivist methodologists, perhaps most forcefully by Simmel who thought that understanding as a method characteristic of the humanities is a form of empathy (in German "Einfuhlung") or re-creation in the mind of the scholar of the mental atmosphere, the thoughts and feelings and motivations of the objects of his study.

It is not only through this psychological twist, however, that understanding may be differentiated from explanation. Understanding is also connected with intentionality in a way explanation is not. One understands the aims and purposes of an agent, the meaning of a sign or symbol, and the significance of a social institution or religious rite. This intentionalistic or, as one could perhaps call it, semantic dimension of understanding has come to play a prominent role in more recent methodological discussion.

If one accepts a fundamental methodological cleavage between the natural sciences and the historical Geisteswissenschaften, the question will immediately arise of where the social and behavioral sciences stand. These sciences were born largely under the influence of a cross pressure of positivist and anti-positivist tendencies in the last century. It is therefore not surprising that they should have become a battleground for the two opposed trends in the philosophy of scientific method.

If understanding and explanation are not the same in the social sciences, and particularly education; if understanding is not just what comes from representing the dependent variables as some function of independent variables, what is understanding? Certainly some cause and effect relationships (or apparent relationships) are a part of understanding. But so also are our experiences, both our ordinary life experiences and professional experiences gained in a context of problem-solving and responsibility. Such experiences contribute ... no, they greatly dominate our understanding of teaching and learning.

Some writers are quick to point out the fallacies of folklore. But if we were able to take a full and impartial inventory of our understandings of pedagogical matters we would find the great bulk of our understandings rooted in personal experience, verified empirically in office and classroom, serving our educational aims both directly and indirectly. That is not to say that folk wisdom is to be preferred to the findings of disciplined inquiry or that folk wisdom should not be challenged. What is true is that folk wisdom has served long and in many instances extremely well in guiding our professional practice. Researchers need not be the missionaries to replace folklore. They should try to help purge it of misunderstanding. And more, they should try to get into its communication circuits and even to build upon its powerful methods of comprehension.

Earlier I claimed that communication beyond the simpler acts of knowledge transmission, can lead to action. Even dialogue that serves little to bring out alternative views, even if ineffective in stating the options, it can be catalytic. Dialogue (communication) is a third component, and as are the other two, a sufficient condition for changes in understanding.

explanation + experience + dialogue → understanding

Moving one more generation back, let us note that experience comes in various degrees of directness, some of it first-hand and repetitive before our own eyes, and some of it remote, only shared through testimony, and some of that hearsay. We may or may not want to be as quick as the courts to honor direct witnessing and to spurn hearsay, but let us postpone talk of the merit of different testimonies. Let us express the variation as

direct experience + vicarious experience → experience

And explanations come to us as a joint effort of holistic and analytic thinking. We devise the more general statements or schemes to cover dissimilar instances. And we locate special instances that test the generality of the hypothesis. Explanations are the product of theory and data, each with a language unlike those of teacher and learner, but a language that permits an expression of *confidence limits.*

formal theory + codified data → *explanation*

Were we to put all the above flow chart equations together, we would have a scheme or map, at least a sketch, but an argument that there are many points of intervention or facilitation. These points should be examined by the researcher desiring to help practitioners solve a problem, perhaps to take a different course of action or to maintain the *status quo*. It does not demonstrate that research can effectively contribute effectively at any point, but it suggests that research may contribute at different points. The pathway from theory and data to explanation -- seen rather simply in many research methods books -- is here seen as tracing many opportunities for the practice-oriented researcher. Attractive alternatives can be seen in ways research might contribute to vicarious experience, improved dialogue, and even to demand.

The educational researcher should continue to hope to produce knowledge in a form that explains things that are happening in education. He should seek knowledge that provides educators with a greater leverage for controlling and improving those events. Some of his efforts should be marshalled toward those aims. But he should not accept the criterion of "knowledge-use" as the sole criterion for his work. One of his purposes is to be of assistance to educators. The products of research are many, and the ways research may influence practice are many. The criteria for good assistance are not the same as the criteria for valid knowledge. Only a few march to a different drummer. But each responds to different voices. There are the voices from without, and the ones from within. What moves a teacher to break an old habit of censuring athletes or to keep a challenged *Candide* on the literature list? What moves an administrator to hire a gifted but controversial project director? These acts are small, but they are the crucial substance of professional practice. They are taken or not taken, depending on what voices are heard. Can research become better heard?

1973

While at the University of Gothenburg, I settled into a conception and ritual for program evaluation, calling it "Responsive Evaluation." Hearing my presentation, a respected Swedish colleague, Ference Marton, commented, "What you're asking for, Bob, is for educational psychology to commit suicide." It was not my intent, nor was I heeded.

Responsive Evaluation

Most of today's plans for the evaluation of educational programs are "preordinate." They rely on prespecification. They emphasize (1) a statement of goals, (2) standardized tests of student performance, (3) value-standards held by the program staff, and (4) a research-type report. It is presumed by some people that these are essential features of any evaluation plan. They are not. There is an important alternative to preordinate evaluation: responsive evaluation.

This is not a new alternative. Responsive evaluation is what people do naturally in evaluating things. They observe and react. They examine the thing, the implications of having it, its worth. What is new is a technology developed around this natural behavior. Much of the error people make in their casual evaluations can be avoided by deliberate readiness, care, replication, and cross-examination. The evaluator does not need to rely on preordinate objectives, experimental controls, or criterion tests to minimize evaluation errors. Poet T. S. Eliot[1] wrote:

Let us go then, you and I,

[1] Elliot, T. S., 1915. The Love Song of J. Alfred Prufrock. *Poetry, VI, III.*

When the evening is spread out against the sky
Like a patient etherized upon a table;
Let us go, through certain half-deserted streets,
The muttering retreats
Of restless nights in one-night cheap hotels
And sawdust restaurants with oyster-shells:
Streets that follow like a tedious argument
Of insidious intent
To lead you to an overwhelming question ...
Oh, do not ask, 'What is it?'
Let us go and make our visit ... "

An evaluator is asked by a client to do an evaluation. They have a program in mind. They have certain audiences in mind. They have certain purposes in mind. But the purposes, audiences and program are likely to change.

An educational evaluation is a "responsive evaluation" if it orients more directly to program activities than to program intents, if it yields information the audiences want, and if value-standards of staff and significant others are taken into account. In these three separate ways an evaluation study can be responsive.

To do a responsive evaluation, evaluators (perhaps a teacher, perhaps a team of specialists) do many things. They make a plan of observations and negotiations. They arrange for various persons to observe and reflect. They prepare brief narratives, portrayals, product displays, graphs, etc. They find out which of these are of value to their audiences.

They gather perceptions and value judgments from significant others, deliberately from those whose points of view differ. They get people to check the quality of the records; e.g. program staff the accuracy of the portrayals, audience members the relevance of the findings, school authorities the practicality of the recommendations. They do much of this informally, iterating, keeping a record of action and reaction.

They choose media accessible to the audiences. They might prepare a final written report, or might not, depending on what they and their clients have agreed on or have need of.

A responsive evaluation performs a service. It is useful to known persons. An evaluation probably will not be useful if the evaluator does not know the interests, problems, and language of his audiences. During an evaluation study, a substantial amount of time may be spent learning about the information wants of clients and other audiences. The responsive evaluator will have a good sense of whom he is working with and for. Anthropologist Charles Frake[2] said:

> *Although my original fieldwork among the Eastern Subanun, a pagan people of the Southern Philippines, was focused on a study of social structure I found it exceedingly difficult to participate in ordinary conversations, without having mastered the use of terminologies in several fields, notably folk botany and folk medicine, in which I initially had only marginal interest.*

Responsive evaluations require planning and structure, but they rely little on formal statements and abstract representations: e.g. flow charts, test norms. Statements of objectives, hypotheses, test batteries, teaching syllabi are of course given primary attention if they are primary components of the instructional or developmental program. Then they are treated not as the basis for evaluation, but as components of the program. These components are to be evaluated just as other components are.

[2] Frake, C. O., 1964. How to ask for a drink in Subanun. *American Anthropologist, 66, 6-2.*

Tests and other data-gathering devices are not ruled out. The choices of these instruments are made as a result of observing the program in action and of interacting with various groups having an interest in the program.

Planning and structure are needed in order to make a proper choice of issues, data, observers, reactors. Besides help with the question "What data to gather?" the evaluator needs special technology for the questions, "How to gather and process the data?" and "How to report the results?"

But always, with a responsive approach, "it depends ... " The methods and the reporting are to be adapted to the circumstances. The initiative for proposing alternatives, and sometimes the choice, rests with the evaluator -- but the selection is based first on what is wanted and later on what is proving useful.

In order for clients to make good choices they need to be able to visualize, to comprehend. They often need examples, illustrations, dialogues, case studies. They need a chance to fit them to their experience.

Evaluating an educational program would be impossible if it were necessary to express all purposes or accomplishments. Fortunately, it is not. It is difficult to be accurate even about a few. It is not necessary to be explicit about aim, scope, or probable cause in order to indicate the program's worth. Explication of intent will usually make the evaluation more useful, but it also increases the danger of misstatement of aim, scope, and probable cause.

To layman and professional alike, evaluation means that someone will report on the program's merits and shortcomings. An evaluator may report that a program is "coherent, stimulating, parochial, and costly." But such simplicity could be misleading. These descriptive terms are value judgment terms. An evaluation has occurred. The validity of these judgments may be strong or weak; their utility may be great or little. But the evaluation was not dependent on a careful specification of the program's goals, activities or accomplishments. In planning and carrying out an evaluation study, the evaluators must decide how far to go beyond the bare-bones ingredients: values and standards. Many·times, they will want to examine goals. Sometimes not. Many times, they will want to provide a portrayal from which audiences may form their own value judgments.

The purposes of the audiences are all important. What would they like to do with the evaluation report? Chances are they do not have any plans for using it. They may doubt that the evaluation study will be of use to them. But charts and products and narratives and portrayals do affect people. With these devices persons become better aware of the program, develop a feeling for its vital forces, a sense of its disappointments and potential troubles. They may be better prepared to act on issues such as a change of enrollment or a reallocation of resources. With the evaluation, they may be better able to protect the program.

Different styles of evaluation will serve different purposes. A highly subjective evaluation may be useful but not be seen as legitimate. Highly specific language, behavioral tasks, and performance scores are considered by some to be more legitimate. In America, however, there is seldom a greater legitimacy than the endorsement of large numbers of audience-significant people. The evaluator may need to discover what legitimacies his audiences (and their audiences) honor. Responsive evaluation includes such inquiry.

Responsive evaluation will be particularly useful during formative evaluation when the staff needs help in monitoring the program, when no one is sure what problems will arise. It will be particularly useful in summative evaluation when audiences want an understanding of a program's activities, its strengths and shortcomings, and when the evaluators feel it their responsibility to provide a vicarious experience.

Preordinate evaluation should be preferred to responsive evaluation when it is important to know if certain goals have been reached, if certain promises have been kept, and when predetermined hypotheses or issues are to be investigated. With greater focus and opportunity for preparation, preordinate measurements made can be expected to be more objective and reliable.

It is wrong to suppose that either a strict preordinate or responsive design can be fixed upon an educational program to evaluate it. As the program has moved in unique and unexpected ways, the evaluation efforts should be adapted to them, drawing from stability and prior experience where possible, stretching to new issues and challenges as needed.

1974

Given the human condition, often it is better to be meangful than precise. How reliable are our portrayals of Education? Do they tell us where we are going?

Navigation toward the Bells of Villingen

My friend Gary, navigator of his destroyer escort, was good at inferences. Every morning and evening at sea he would infer his ship’s position, partly by measuring the angles of elevation of the stars above the horizon.

Just after sunset he would bring his sextant to the bridge. "Damn clouds," he would mumble, then, "Ah, there's Vega," and on about the business of taking “his fix.” His habits no longer needed the voice that said, "The ship's position is somewhere on an imaginary circle centered at the point on the earth's surface at which at this moment Vega is directly overhead. The size of the circle is determinable from the angle of elevation." For every moment of every day, Gary’s tables told him, for the elevation of that star, where the center point was, and he could then plot an arc of possible locations on his map.

Of course, it wasn't sufficient to know that his ship was located on one particular but very large circle on the face of the earth. But when he got the elevation of Dubhe, that Big Dipper star to the north, he could draw a second circle that intersected the first.

The two points of intersection were usually thousands of miles apart so, on a slow ship like his, and mine, which couldn't possibly be in an ocean different from yesterday, he could infer where we were from two star-sightings. But, for greater accuracy, he would take another sighting. The arcs of the circles, so large as to look like straight lines, would appear as three lines intersecting on his map.

Neither Gary’s chronometer, nor elevations nor calculations were perfectly accurate, so the lines (circles) did not intersect in just one point. He would take six stars, if the clouds would let him, and then, finalize his fix with a dot in the middle of the intersections. His book called this process of approximation, "triangulation," though I don’t recall hearing him use the word.

Two Situations

I believe we who measure education make the same sort of inferences. "Where *are* we?" is not an uncommon question in the teachers' lounge, boardroom, and commissioner's office. The task is to observe the surroundings, to calculate progress, to gather whatever direct and indirect evidence we can, and to map out the present circumstance.

Our task is more difficult because education is a space of a hundred dimensions rather than the *three*-dimensional surface of a sphere and because our maps, tables and sextants are not nearly so refined -- but perhaps a less crucial task, because there is so little chance that our measurements will crash us on the rocks.

The initial task of the measurer is observation. We find ourselves in two situations: (1) where one or a few attributes have been specified to be measured and (2) where we count on our experience to recognize our accomplishment. We may have been commissioned to measure readiness to learn, the simplicity of the syntax, or the vindictiveness of the voters. Or perhaps we were commissioned to observe a group of children, a science curriculum, or vocational-education legislation. These two are quite different commissions, (1) the measurement: closed and (2) the observational: open.

In carrying out a measurement assignment or contract, or in doing much ordinary educator work we often are to move from (1) to (2), pre-specifying the sightings or signs of progress along the way. Sometimes we just head for (2), guessing (dead reckoning), observing, counting on past experience to tell us the success of our work. Let us keep in mind these two measurement situations, the closed, with pre-specified criteria of progress, and the open, counting on our own experience to recognize accomplishment. (Someday we might be calling the situations: the quantitative and the qualitative.)

In both situations we are concerned about the reliability of the description. At sea, the compass points have been fixed, the attributes of our location are specified, the measurement situation is closed. The Captain looks over Gary's shoulder, noting not only whether or not we are on the course intended but also the size of the navigator's triangles. If the arcs do not converge almost to a point (maybe the horizon was hazy, maybe he was thinking of Yleen), the fix is not reliable.

In education we want to know the children's understanding of magnetism or the history of the segregated classroom. We cannot plot these on two-dimensional maps. The dimensions are many, and they fade in and out of importance. Still, we want our observations, our discernings of presence, difference, and uniqueness to be reliable. What should we borrow from the navigator? How can we communicate holistic insights?[3]

Accuracy

Observations are reliable (by my definition) if their record can be relied on as properly descriptive. Essentially this means that they must be accurate representations. The triangles must be small, the standard error of measurement must be small, for the description to be reliable. To be accurate is to indicate a proper amount but also a proper shape, proper

[3] Rhyne, R. F., 1971. Fields Within Fields -- Within Fields, *Evaluation Practice*, 563-76.

gradations, a proper context. It may be a simple accuracy. The fullness of a cylinder of fluid can be a reliable measure of rainfall, or of the navigator's abstinence, or of the temperature of the classroom. Or it may be a complex accuracy. A portfolio of quotations, correspondence, and sketches can be a reliable measure of an art teacher's qualifications or the testimony in a lawsuit. Accuracy includes the idea that the attribute or object or action is properly represented. Statements or scores are reliable if they can be relied on as suitable descriptors.[4]

To be reliable may or may not mean to be free of bias. Bias is a systematic distortion of the description -- usually unintentional, caused sometimes by a faulty instrument but more often by human priority or habit. One admissions test I know is biased -- it asks more difficult questions than are appropriate. One observer I know is biased -- he sees all state education agencies as bumbling, anachronistic. There are those ethnic, political, religious, artistic, athletic, methodological biases -- some shameful, some bothersome, some benign. Any bias -- especially when paraded as virtue -- can be built up to cause personal injury, embarrassment, or loss of rights and privilege.

Measurement is as laden with bias as any communication. Safe-guards are needed, not so much against bias as against bias that injures or confuses.

You cannot: draw a real line between good bias and bad. Bias comes with caring, with probing, with interpreting. To eliminate bias would be to eliminate scrutiny. What Social Scientist Paul Diesing said for the scientist is true for the teacher

[4] When we are using measurement to actually describe an object or situation, we will speak of the reliability of the measurement. When we are using a measurement as an indicator, such as to indicate a classification for the object or to choose a treatment for it, then we will speak of the validity of the measurement. Each time we consider the value of the measurement we will need to consider its validity for the action to take. Then the measurement will be more than mere representation of the object.

or technician as well and for anyone who measures. Deising[5] said:

> *Every scientist must perceive and interpret his subject matter from some standpoint and thereby bias his conclusions. Also every scientist must be active with his subject matter in some fashion and must therefore change it as he studies it.*

All reports are based on certain preferences, certain beliefs, as to what is worth reporting. All instruments reflect commitments of their makers. The preferences and commitments need scrutiny to face the challenge of the purposes and jeopardies of the immediate situation. Many will be found biased, but acceptable, not causing an objectionable unreliability in the reporting.

Reliability is not an impersonal characteristic belonging alone to technology. To be reliable, a measurement must be capable of being relied on by people as a suitable representation of the attribute or object. The criterion for reliability is effectiveness of representation; and the effectiveness is something judged by persons, by users. We may not allow a single individual to decide, but ultimately the judgments of' a collection of people will decide.

To be an effective representation the measurement must have a certain accuracy. Accuracy alone does not make the measurement reliable. The effectiveness of representation for choosing what to do makes it valid.

Consistency

In defining reliability most measurements experts refer

[5] Diesing, P., 1992. *How does social science work*? University of Pittsburgh Press, p. 279.

to the concept of consistency. "A reliable teacher is one who consistently does her work. A reliable test is one that, on repeated testings, would give the student consistent scores.” They point out two different consistencies, consistency over time and consistency of different observers (or instruments).

It is troublesome to apply the consistency-over-time definition when the attribute or object is changing. Students learn and grow weary; teachers have ranging purposes -- which make it almost impossible to know whether or not repeated measures are accurate.

In the open-observation situation the observer may visit the same scene repeatedly and make reliable reports, and each one is different. Good observers will take a second look, and a third, and a sixth; they will count things twice and look for confirmation. We measurements people should look for consistency over time; we should be hesitant about making it an integral part of our concept of reliability. When you *have* consistency, you have a small argument for reliability -- but when you cannot show consistency, you have a small argument for unreliability -- but in either case the argument cannot stand alone.

If the phenomena are not constant, observations will not be consistent. I thought:

I watched the diamond turning slowly in her fingers. The sparks were lightning quick, not anticipated nor followed by glow. I found no way to predict their burst and color. Somehow the cycle did not repeat itself.

I watched ten-year-old Jacob playing Twenty Questions. He sometimes asked a question which had worked before. Seldom did I think it the best for the moment. But his questions changed, got better. I tried to anticipate what he would ask, but I could not. Such a simple game, such a simple learning situation, defying my efforts to sketch a simple scheme for *describing the regularity or progression of his questioning.*

I watched myself watching diamonds and children's questions and I wondered how my observations could be trusted.

There is the problem of looking at the right things, of course, and there is the problem of seeing what you are looking at, and there is the problem of telling what you saw. You have to do all three to make a reliable record.

When I did not find consistency-over-time watching diamonds, I was easily persuaded that the attribute had varied or that the action was irregular. Consistency-over-time is in fact an inadequate test for reliability of educational observation in the open-measurement situation. As the situation closes (to a situation like that of the navigator’s) -- as one or two dimensions or attributes become the focus of the measurement -- and the movement is slow, consistency-over-time may help us test for reliability.

Consistency-over-observers is more useful, but here too there is trouble. Different observers of the same scene make reliable reports in the sense that it is found that they can all be relied upon to describe the same scene, but all reports are different. Of course, no one describes the whole scene, and most concur on some aspects; but each result comes out different.

For good measuring we will not only tolerate inconsistent observations but will solicit them. Multiple observers and somewhat parallel instruments will often be used to provide a reliable report on education. The more independent the observers are of each other, and the more distinctively different the instruments are, the more chance that right things will be observed, that what is seen will be seen accurately and that what is reported will be reported effectively. Agreement may stem from a joint blindness or common bias, but agreement remains worth looking for. An important confirmation -- though seldom full authentication -- comes from consistency-over-observers. The reader of the report bears some of the responsibility for resolving the inconsistency observed.

Measuring Errors of Measurement

Reliability itself is an attribute, subject to measurement. Sometimes we can make a quantitative statement about the amount of reliability, perhaps calling it a coefficient. Sometimes we will want to observe and describe reliability in non-statistical ways. For the closed-measurement situation the statistical ways of measuring reliability have been well formulated by measurements experts such as Julian Stanley[6] and in textbooks such as that by Jim Popham[7]

Good technicians and methodologists in any field want to be specific about how much confirmation, how much reliability, particular measurements have. They want to be specific about the error. A vast technology has developed for the educational test industry, as for many other industries, to specify the degree of measurement inaccuracy. A chemist specifies the weight as 11.03 grams, noting ±.01 gram to indicate the expected or tolerable margin of error. A test manual indicates the standard error of individual achievement scores to be 4 points, indicating we can expect that usually a "true score" for the student would be no more than 4 points higher or 4 points lower than the one he gets.

Psychometric statements of expected error are usually generated by considering the variability of multiple

[6] Stanley, J., 1971. Reliability. In R. L. Thorndike (Ed.), (1971). *Educational Measurement* (2nd ed., pp. 356–442). Washington, DC: American Council on Education.

[7] Popham, W. J., 2006. *Assessment for Educational Leaders* (2006), Allyn & Bacon.

observations (multiple scores, multiple items, etc.) of the students. The rationale for a standard error of measurement is based upon whether or not the variability of one student's scores would be small compared to the variability of the scores of a population of students.

In *Educational Measurement*, Julian Stanley (See Thorndike, R.,1971, p. 359[8]) wrote:

> *The evaluation of the reliability of any measure reduces to a determination of how much of the variation in the set of scores is due to certain systematic differences among the individuals in the group and how much to other sources of variation that are considered, for particular purposes, errors of measurement.*

Stanley was talking about the definition of reliability widely accepted in a measurements community that has reduced its field to those situations when observation (testing) is standardized on preselected attributes (the closed situation) and where variability of individuals (objects, actions) has been accepted as the proper basis for describing the single case. There are other measurements communities.

When an observer measures a child's difficulty in learning to read or a superintendent's difficulty in protecting a literature program, no reference group of students or superintendents is apparent or essential. The question is not whether or not to use statistical language but whether or not to use the conceptualization of error favored by certain statisticians. The measurement of an educational situation and the measurement of the error in measuring that situation go

[8] Thorndike, R., 1971. *Educational Measurement*. Washington, DC: American Council on Education.

hand in hand. On those occasions when people are not conceptualizing the individual as one among many, when they prefer a direct description of the case, a population concept of reliability is inappropriate. The audiences want measurements that fit their conceptualization of foregrounds and backgrounds, ones they can rely on as fitting their own ways of observing and describing. What has been measured should be clear, the representation accurate and comprehensible, with indication of reliability built into the description.

Comprehensibility

Accuracy and comprehensibility sometimes get in each other's way. I once visited Villingen, a high-walled Black Forest village. The church bells of Villingen were a special treat and their story too. First one church would toll the hour, then another until finally perhaps a sixth. Obviously, if they tolled together, it would be difficult to count the hours. But if each waited its turn, some would be inaccurate. My host, Christa Loercher told me that Villingen's choice had been for meaning over accuracy.

For the measurer in the closed situation, with attributes pre-specified and especially with direct measurement of attributes possible, accuracy and comprehensibility often are compatible. For the measurer in an open situation, there is always the search for compromise between the representation that can be accurately obtained and the representation that can be comprehended as standing fully for the object itself.

How comprehensible it is depends on the reader and audience. A reader or audience is a something needing to be known -- if measures are to be reliable. What will they comprehend? What discriminates? What uniquenesses will they recognize?

The measurer is planting a measurement seedling within a forest of existing experience. Can it survive a drought of disinterest, the storm of disbelief? It is not the measurer's job to overturn disinterest or disbelief -- measurement is not evangelism -- but it is the measurer's job to provide a representation that has a chance of survival in minds grown to shelter certain concepts and not others. Something cannot be relied on if it lacks meaning for the potential user.

A few paragraphs back I downplayed the existing statistical definitions of reliability, those that placed an individual (was it student or superintendent) within a population or reference group. Now I entertain a statistical definition, one where new measurement joins a shelf-full of methodological ideas, images, icons, and indices. If the new lies outside the range of the old, it probably will not be assimilated. If it is too vague, it cannot be housed. If its variability (across perhaps six repeated measurements) is large compared to the variability known to the existing audience, then it is not eligible to be considered a reliable descriptor. The experience and concepts of any one audience will not be available in the data bank or in the archives, but they can be sampled.

It has not been unusual for a measurements specialist to say, ''Here are the test scores and here is the technical manual. It is not my responsibility to make them useful to you." Still, for the last fifty years, the teaching in measurements courses has been quite good. Standardized test scores were taught as "individual-differences-reliable" and student were taught what this meant. And yet many students did not comprehend it. Many teachers and administrators continue to not use test scores appropriately.[9] The burden of audience comprehension is not the measurer's alone, but he or she is first to bear it. So perhaps it is time for a different course.

[9] Hastings, J. T., Runkel, F. J., and Damrin, D. E., 1961. Effects on use of tests by teachers trained in a summer institute. Cooperative Research Project #702. Urbana, IL: Bureau of Educational Research, University of Illinois.

What we saw in Navigator Gary was not a dependence on simple longitudes and latitudes but a reliance on multiple reference points and successive calculations and interpretations to determine "where we are."

Part of the solution is the choice of language, both for expressing the observation and for expressing its susceptibility to error. Researchers, philosophers, teachers, and taxpayers do not share a common language. The measurer should choose separate concepts, vantage points, instruments, and coefficients to fit the thought-pace of the audience.

Part of the solution is to resist the pressures to over-standardize the technique of measurement and to resist apologizing for observation. We should choose what will be measured not because it lends itself to scaling and variance analysis but because it measures what needs to be described and comprehended. There will be a trade-off between using an existing scale or checklist and custom-building a new one. The advantages of a well-established procedure are many; the chances of a here-and-now procedure being as reliable are small. Still the existing measure may be too general or too specific, or too much of something else, for the present purposes. In choosing new or old, the measurer can give more attention to audience comprehension.

Part of the solution is to look for confirmation in a broader field, not just in repeated measures or parallel instruments. When Gene Webb and Donald Campbell[10] wrote of using "unobtrusive measures" in sociological research, they spoke too of *triangulation*.

It is through triangulation of data procured from different measurement [approaches] that the investigator can

[10] Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L., 1966. *Unobtrusive Measures,* Thousand Oaks, CA.

most effectively strip of plausibility rival explanations ... The usual procedural question asked is: Which of the several available data collection methods will be best for my research problem? We suggest the alternative question: Which set of methods will be best? -- with "best" defined as a series which provides data to test the most significant threats to [an interpretation] with a reasonable expenditure of resources.

Webb and Campbell spoke of broadly distinct and potentially contradictory manifestations of the same phenomenon. The nose prints on the glass, the wear of floor tile in front of the exhibit, the count of time spent by a sample of visitors, the choice of subjects in a composition on "Our day at the Museum" may or may not triangulate. The measurer too should seek multiple independent entries for confirmation of his fix.

Efforts to Disconfirm

Gary Joselyn relied not on the stars alone. In sight of land, or in radar reach, the azimuths and distances made plotting easy. To us far at sea, Loran radio beacons sent vectors. The depth of the water confirmed or disconfirmed a plot. Using gyrocompass, ship's speed and time, by "dead reckoning" the navigator would project forward from an earlier plot. (Confused we sometimes were, but almost never without opportunity to confirm.)

One more thing about the navigator's routine, perhaps instructive to the measurer: In getting his star fix, the navigator in fact started with an estimated plot and, with elevations and calculations, corrected it. I wonder if Gary thought about it -- the powerful methodological difference between creating-from-scratch and correcting-and-shaping. When the task is terribly complex, the only way to achieve success may be to set up a preliminary result and improve upon it. "Progressive focusing" Malcolm Parlett and David Hamilton[11] called it. The strategy of the measurer in such a case should be to invest part of his effort in correcting first measurements, in deliberately seeking challenge and disconfirmation. The measurement is fixed when no evidence is found to indicate that further correction is needed.

What an audience needs to know, in addition to the measurement, is how the observer looked for disconfirmation and what he found. On many occasions, especially in the “open” measurement situation, the data will not reduce to a single fix -- but no further correction is practical or perhaps seems possible:

The school psychologist administered the WISC, found Nancy's IQ then to be 75; Nancy reads very slowly; the kids say she is a "dumb-dumb"; her older brothers were slow students; she never reads newspapers or magazines at the news counter where she works; yet she makes change quickly and accurately.

The measurer seeks the best inference, reports his or her fix, and lists the confirming and disconfirming findings. The test of strength of the reliability of an educational measurement lies less in its replicability but more in its resistance to disconfirmation, in its continuing effectiveness of representation, under challenge.

[11] Parlett, M. and Hamilton, D., 1977. Evaluation as illumination: A new approach to the study of innovatory programmes. In Hamilton D., Jenkins D., King C., MacDonald B., Parlett M. (eds), *Beyond the Numbers Game.* London: Macmillan. pp 6-22.

1975

During my sabbatical year, 1973-74 I was on SAFARI. An English safari, not African. I worked as a measurements consultant on a summative evaluation project with Barry MacDonald and Rob Walker. SAFARI was a Ford Foundation project to follow-up four completed British curriculum development projects. In particular, it was to look at the legacy of one of the four. More broadly, it was to consider the legacy of the course-content education-development movement. The projects had lived long enough to feel they had a legacy. Along the way the general question came up as to how much to charge the evaluator with examining the administrative and political conflicts therein.

SAFARI and Legacies

The acronym SAFARI stood for "success and failure and recent innovations." It was a summative evaluation project attempting to contribute to better higher education by studying curriculum-building. A key aspect of the study was a search for the criteria by which projects, whole projects -- in an ultimate sense -- are evaluated.

Different persons choose different criteria for summative evaluation. There is often little correspondence, for example, between the criteria set by an evaluator affiliated with the project and these criteria set by a teacher or administrator in a field trial school. There are "multiple realities" of success and failure. A measurements specialist is likely to give focus to student· achievement or to enumerated exchanges in the classroom, or to attitude scale responses of teachers. An educator -- when evaluating -- is more inclined to give focus to singular incidents, especially those of personal immersion and program interruption, those bringing pleasure and embarrassment. The one more often chooses attributes, the other episodes.

Their methods differ, but also the essence of what is important to them. The measurement person's emphasis is usually on the formal characteristics of the project, description -- often discipline-based description -- of process and product variables. The participating educator's attention is often on the project's compatibility with ordinary school operations. Of course, it is simplistic to think of just two sets of criteria, there are others. In the following paragraphs I will identify several classes of success and failure manifestations of innovative projects.

The project evaluator needs to look at what is wanted. At the top of most WANTED lists, of course, are changes in student learning, teaching practice, or administrative arrangement. Some of the anticipated instrumental effects are likely to be featured in the original project 'proposal.' Other pay-offs are anticipated but not mentioned. Still others are unanticipated. We look also at actual accomplishments. “They innovated” to change things. Often the innovation is a tool to fix something, and the evaluator looks to see if it got fixed.

It is often assumed the worth of the change varies with the durability of the change. Not only “Is it fixed?” but 'Does it stay fixed? Does the fix take root?” The first question Barry MacDonald raised in the SAFARI proposal was: “How do new educational ideas survive?” Survival, as a term associated with life, an interval of time terminated by death. The question arises, “How long should an educational innovation survive?” Three years? Ten years? Of course, it can be a failure though it lasts a century. And it can be a success if it survives until Christmas. Durability is not always a good indicator of success, but it needs to be considered.

Ford Foundation officials and many investors in innovation have spoken of a “multiplier effect.” They feel that the immediate instrumental effects of the innovation cannot justify the project costs. Justification is sought in multiple usage, diffusion, adoption in more and more places, affecting more and more people. Counting and mapping usage is an important activity of the evaluator of innovation. Project success or failure remains dependent, of course, on the quality of the usage.

Some projects do not survive, but illuminate the way for other efforts. Some projects do not take root, yet succeed as a stepping-stone, enabling movement to a more distant enterprise. The message on the first telegraph wire was said to be, 'What hath God wrought?' The answer, certainly, is more than four words, and more than Western Union. The telegraph did more to advance rail transportation than it did to advance personal communication.

In Education, the programmed instruction activities of Pittsburgh's Learning Research and Development Center did not survive but they were the foundation for Individually Prescribed Instruction. The Illinois Area Service Centers could not have become the resource they are without the experience of the now defunct Demonstration Centers of the Illinois Program for Gifted Children. Sometimes enablement is a more important outcome than 'taking root'.

Instrumental effects are important indicators of success but often difficult to measure. “Changes aren't apparent. The problem isn't fixed.” Even then the innovation may be partially successful in terms of *habitat effects*. These are what the innovation does to improve the environment, to make the school (or whatever) a better place to teach, to learn, to live. What instrumental effects these purchases will have are not only seldom measured, they are sometimes beyond speculation. Often it is assumed that there will be some, sometimes not. The justification for the innovation may even be first-of-all that it improves the milieu. We would not deny that an innovation occurs partly because it is time for a change. We have an aesthetic sense of what the school should be. A mirror is used both for adding eye shadow and soul seeking, and in neither sense is a luxury. Innovations sometimes improve the looks and feel of thing.

Adjusting for Inevitables

There is a particular criterion of project success -- one that falls into the habitat class, I think -- one that deserves more attention than it gets. We live in an expansionist, acquisitive society. Like the "Class of 1973" and the "Class of 1974," each successive wave of officials will leave its mark. It has a quota to fill. Funds have been budgeted. There will be purchases of new educational goods. Given that there will be purchases, the evaluation of the innovative project may be largely dependent upon whether or not the innovators acted in good faith and used good judgment. The measure of success then is not. just what they changed, but how well they used their opportunity.

One of the four SAFARl projects was created partly because in England the school-leaving age was increased a year. Thus, the new pupil population included a large number of older adolescent youngsters. "Something had to be done!" Was this project an appropriate response, given that there would be a Schools Council response? Efforts should be considered at least partially successful if resources sure to be spent were used to purchase what seemed at the time a reasonable accommodation to stress and change.

Of course, it may not be politic to state in the report that the resources were sure to be spent. This all started bothering me when an acquaintance told me about a project in a rural "developing" country. The backwoods teachers were not doing a good job of giving the youngsters what in Britain and the U.S.

is considered a basic education. The best remedy that the project came up with was an educational television supplement to regular teaching. Many did not expect it to make a difference, but it seemed worth a try. Now it's evaluation time, and they are not going to find an improvement in what the learners have learned. Was the project a failure?

A sponsor such as the Ford Foundation or the Schools Council wants to spend its funds where they will do the most good. It is important for the funder -- and for us -- to find out if they got their money's worth. That is hard to do.

And it is even harder to do if we recognize that the funding agency, private or governmental, is committed to spend certain moneys. Usually, it is an agency to spend money, not one to decide whether or not to spend. Its alternatives turn out to be surprisingly few. There are too few good proposals, and each one promising more than it can deliver. The agency is concerned about image and political pressures, and rules out certain proposals, and invites a few others. The result is that there are moneys to be spent and not many acceptable good-risk alternatives available. The importance of the cost factor diminishes considerably, but how much is not clear.

The agencies behave much like individual persons. Even in a time of tight budgets people respond to large and perplexing problems by shopping around for new solutions. New programs will be started, though there is little research or experience to· recommend them. Maybe they should not but they will. The value of the project should not be confounded by the merit of our propensity for creating new projects. The propensity should be evaluated separately. Given the propensity, the innovative project should be evaluated partly in terms of whether it was one of the more reasonable efforts to remedy a problem or improve the setting in which the problem was being worked on. The project may be seen as a proper undertaking, indeed even a partial success, though it fails to solve the problem or bring about the hoped-for changes.

I think we need this broader definition of legacy. If evaluation were to be based entirely on the quality of process and product, with due attention to compatibility, it is likely that in the long run evaluation reports would discourage efforts to solve problems and improve the school environment. Certainly not all innovative efforts deserve praise. Extravagance and whimsy should be kept in check, but evaluation should not become the agent of those who believe that educational problems -- even any one of them -- are beyond remedy.

In looking at the legacies of projects, as we did in SAFARI, we who would try to measure education should not fail to raise the question of whether or not each project mobilized some of the best resources available and mounted a reasonable attack upon the problems of' the time. This is more than a check into the quality of the process of the project. It is a check into the contribution to an ethic of enquiry and improvement

Tracking Mistakes

On reaching this point, and reconnoitering as good safari-goers would, I wonder if I am pressing too hard to find the Good in a curriculum project. I recognize that formative evaluators, especially internally-appointed formative evaluators, are usually empathic, virtue-seeking, and, protective. As Michael Scriven[1] warned, they are co-opted.

But I recognize also that after a project has terminated, after the ego-involvement has diminished and dispersed, many participants and evaluators find fault more than accolade. A few

[1] Scriven, M., 1967. The methodology of evaluation. In Stake, R.E., editor, *Perspectives of curriculum evaluation*. Chicago: Rand McNally.

'of the managers struggle to protect their credentials.[2] But there is little effort to keep looking for evidence of Good.

Mistakes are important. We can learn from mistakes. In organizing SAFARI, Barry MacDonald was not necessarily critical in promising to look especially for dilemmas and crises in the four terminated projects. Probably the best direction to take for improving curricula is through' the study of past shortcomings.

But the track divides here. Emphasizing past mistakes serves the purpose of understanding curriculum development, and can be access to the study of broader educational issues. But a focus on mistakes imbalances the overall understanding of success and failure of the individual project. A choice of destinations must be made: it is to understand all about the particular project or to understand a little more about the development process.

Our SAFARI chose the track to the left, the wider view. With MacDonald, Rob Walker said:[3]

> *... it is clear that SAFARI involves more than a comparative analysis of ... projects in an attempt to distribute labels of 'success' and 'failure' on a basis of merit. For SAFARI, the comparative study of the projects selected must be a means of reaching out beyond the formal bounds of each project into the education system at large ...*

Does that mean that SAFARI should not have concerned itself with legacies? No. Choosing to seek understandings of the general problems of curricula and education did not relieve SAFARI of the burden of examining the success and failure of each project, but of the burden of providing a full and audited

[2] House, E. R., 1973. *School evaluation: The politics and process*. McCutchan.

[3] MacDonald, B. and Walker, R., 1975. Case study and the social philosophy of educational research, *Cambridge Journal of Education, 5, 2-11.*

accounting. Mistakes, crises and dilemmas are not understood in the absence of success and failure information. Each problem a project staff faced is better understood by considering the legacy of the project. One thing that the choice does mean is that not as much time is available for examining individual project success.

Too much attention can be paid to the legacy, particularly to the instrumental (pay-off) effects. We often hear people say that they would prefer 'to let History be the judge.' I object to that. Long term consequences should not be given short shrift, but the most important standards for judging the worth of a project usually are contemporary standards. A summative evaluation should not overlook the criteria and perspectives of the project's original constituency, supposing that subsequent observers have better grounds for evaluation. If a project is to be properly appraised, it should be considered as a creature of its own time and place, not of a later time and changed place.

The evaluator searches for a legacy and delivers -- at least for his own musing -- a eulogy. He selects a few things to remember. And in a sense his report is a eulogy. By the words of his reports will be known a little more of what once was. By his words will some others set values for what now is. Eulogizing is a worthy tradition. The past is too easily damned. A summative evaluator needs to be reminded.

Part of the reason for the admonition -- and the admonition of this whole essay -- is that in matters of evaluation we become more critical over time. What is genuinely an exciting idea at the time may be later called a fad. “Fadism” is a cliche. Fad is pejorative, inviting one to cease and dismiss serious evaluation. An evaluator should call a spade a spade, but not always a fad a fad, for it may stereotype an innovative project unfairly.

The legacy can be honored, even while every fault is tallied. The evaluator need not deny the transitoriness of the project, though it shares Keats' epitaph: HERE LIES ONE WHOSE NAME WAS WRIT IN WATER. Nor need not ignore broken promises. Evaluators may tell of insignificance. But they honor the legacy by telling of the project as a happening, organic, inter-active, real and immediate, one that made its place a different place, as well as that it made something better.

SAFARI took the track to the left, to contribute to general understanding about curriculum development by studying four projects. Its aim was not to eulogize. Its obligation was not to provide an accounting of costs and benefits. But it could not relate particular dilemmas and crises to education elsewhere without conceptualizing the legacy of each. It probably could not talk about “how new educational ideas survive” if it failed to learn the character of its four projects -- how each was or was not a response, a stepping-stone, an improvement in habitat, and a spark to the spirit of inquiry.

1975

As I closed the first decade of being a program evaluation specialist guy, my views were becoming clearer and less conventional and further distanced from my doctoral training. I passed this following memorandum to my colleagues, Tom Hastings, Jack Easley, Ernie House, Gordon Hoke and others, including CIRCE visitors Ulf Lundgren, Helen Simons, and Diane Reinhard.

The Emergency School Aid Act (ESAA) was the only federal educational research initiative under President Richard Nixon, who opposed federal directing of school districts on how to conduct education and approved this experiment at giving districts massive funding to create their own innovations under a blanket of evaluation.

Resigning from the ESAA-SDC Meta-Evaluation Panel

To: CIRCE staff
From: Bob Stake

On January 8, I attended a meeting of the ESAA-SDC Meta-Evaluation Panel to explain my reasons for resigning. SDC (Systems Development Corporation, Santa Monica) holds a large contract with the U.S. Office of Education for the evaluation of ESAA, a federal act to support schools having racial isolation and desegregation problems, particularly with their instructional programs.

The act is in its second year. About 176 schools in 55 school districts are involved. John Evans of USOE had decided to design it as an experimental program, using the 1 % of total funding to pay for the conduct of an experiment. Mike Wargo of OE is in charge of the design and operation, and John Coulson of SDC is in charge of the evaluation contract. The grand plan includes a meta-evaluation (evaluation of the evaluation) to be directed by Michael Scriven.

The experimental treatment here is "money," payments to the participating districts. Each district has identified its eligible schools. OE paired them and randomly named one “experimental” and one “control” and got promises from each superintendent that ESAA money would be spent on the experimental schools and no over-budget money on the control schools. (It was a promise not kept.)

The main dependent variables are basic-skills achievement, racial balance, and perception of discrimination. The California Achievement Tests were selected to pretest and posttest reading and mathematics achievement in Grades 3, 4, 5, 10, 11, and 12. A "school climate" questionnaire was developed by people from minority-group organizations to measure perceptions. Hundreds of additional characteristics about setting, curriculum, staff, funding, etc. are being obtained from students and district staffs as well. Further, there is to be a ten-day observation in 30 schools this winter by people from SDC to provide case-study information.

The analyses are primarily statistical analyses: regression analysis, discriminant analysis, and others specially designed for the circumstances by Mel Novick and David Wiley. It is important to note that these design people, including Wargo, Coulson and Scriven are currently among the most knowledgeable methodologists in educational research.

Eighteen months ago, Michael Scriven asked me to join him, Dan Stufflebeam and Gene Glass as a Meta-Evaluation Panel, a group of advisers to keep tab on developments and to provide formative evaluation feedback to OE and SDC. Demurring, I expressed my opposition to Evans' idea of using a federal-support program as a randomized treatment experiment, believing that the purposes of research (information gathering) to be too much in conflict with the purposes of ESAA to justify the demands made on participating districts. Further, I expressed the belief that large randomized-treatment research design and multivariate data analysis both had poor records of providing information useful to policy setters and practitioners.

I told our friend Michael that I did not want to be on the Panel, partly because I had too much to do, but he got me to sign a contract for consultation, prepare a statement of opposition, review some of the materials, and to be recognized officially as part of the Panel. I did only the first and fourth of these, not the work part.

Recently I was dismayed to see myself listed as a panelist on an SDC brochure describing the evaluation activities. A friend pointed out that I was inconsistent, advising some researchers to avoid using standardized tests as criterion variables in curriculum evaluation but condoning it elsewhere. But mostly because the USOE-SDC design called for such a simplistic use of tests within a prestigious and costly analysis, I resigned from Michael’s panel.

At Michael's request, I met with the panel to explain myself. I spoke of my objections to the experimental approach and to reliance on multivariate analysis in a school support program. Just six years after CIRCE's reviewing and closing its Illinois State Testing Program, I objected to measurement of the effectiveness of basic-skills teaching using standardized (either norm-referenced or criterion referenced) tests.

How good is the case against such research design and test use? I do not know. I find my support in writings by Lee

Cronbach,[1] Donald Campbell,[2] Egon Guba,[3] Gene Glass,[4] David Hamilton[5] and others. There is apprehension about the increased use of these tests as criteria of educational accomplishment. Maybe we should be spending more time evaluating those methods of measurement and the analyses considered "most respectable."

With reluctance I withdrew. Those are good people with whom to be working. They are discussing issues of interest to us all, such as the possible need for re-standardizing achievement tests for special populations, the anonymity of personal and school information, the methods of measuring discrimination in the schools, and the role of advisers in such an investigation. But I felt that, in the matter of the brochure, I was misrepresenting myself -- and possibly I was being used -- so I did.

[1] Cronbach, L. J., 1980. Validity on parole: How can we go straight? In *Proceedings of the 1979 ETS Invitational Conference*, pp 99-108. San Francisco: Jossey Bass.

[2] Campbell, D. T. and Stanley, J. C., 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago IL: Rand McNally.

[3] Guba, E. G., 1978. *Toward a Methodology of Naturalistic Inquiry in Educational Evaluation.* Monograph #8. UCLA: Center for the Study of Evaluation.

[4] Glass, Gene V., 1979. "Policy for the unpredictable (uncertainty research and policy)." *Educational Researcher* 8, 9, 12-14.

[5] Hamilton, D., 1980. Generalization in the educational sciences: Problems and purposes. In Thomas Popkewitz and Robert Tabachnick, editors, *The Study of Schooling: Field-based Methodologies in Educational Research.* New York: Praeger.

1975

A school district in Michigan dealing with some testing initiatives appointed a review panel consisting of Jan, a local high school teacher, Neil, a staff member of the National Education Association, and myself as a measurement specialist from the University of Illinois. Our dialogue, on-again, off-again, went something like this:

Teachers versus State Testing

Bob: How many people have given us testimony?

Neil: 47, by my count.

Jan: And at least 20 more tomorrow.

Neil: We haven't heard from many teachers yet.

Jan: Most of tomorrow's will be teachers.

Bob: Certainly not everyone supports the Teachers' Association position that too much time is being spent on testing.

Neil: But the majority do.

Jan: I wonder if we can pay much attention to percentages. The people who have come in are not a good representation of the district. I worry about this testimony being discredited because it has been taken from volunteers. Of the 20 or so who were not staff members, most have been mothers of school children.

Bob: I don't think we should worry about representativeness. We are trying to gather information and perspectives. We shouldn't "average" them. We should analyze them and report our interpretations to the Association.

Neil: And to the Board.

Jan: I didn't expect a research specialist to talk that way.

Bob: Most make more fuss about it than I do. In some studies, of course, it is very important. But the point here is that we won't claim representativeness. We'll report that we heard from the Superintendent; one board member; several of the administrative staff, including the Director of Testing; and so on. Some may choose to discredit our findings because we did not hear from students, or from the Michigan Assessment people, or from a sample of taxpayers.

Neil: That reminds me, do we know whether the State program is going to require even more testing time in the future?

Jan: It was expected, but the State Board has told them not to expand. But I think they intend to continue to involve as many volunteer districts as they can in field trials of the criterion-referenced tests they've been developing.

Neil: This District won't be volunteering, will it?

Jan: After opposing the State Program for five years, hardly. The reason the District has its own Objective-Referenced Testing is that the teachers recognized immediately that the objectives of Michigan Assessment did not fit this district. So we drew up our own. I thought we had an agreement with the State Superintendent that would excuse us from the state program once we got our own going, but I guess not. We had task forces in nine subject-matter areas. So far we have only developed tests for the communications skills, but the implication for the future is that tests would be developed in all nine areas.

Bob: Then would any time at all be left for teaching?

Neil: But these are the teachers' tests. I think that we can assume they will fit in with and contribute to the teaching.

Bob: Some of the teachers testified today that the Objective-Referenced Tests were not very good.

Neil: Oh, well, it's the first year. They'll get better.

Jan: But to keep up with what the teachers are teaching, the tests will have to continue to be revised. Can the District afford to give that much teacher time to testing?

Neil: Right. It's not just the classroom time that is being used.

Bob: One teacher estimated the annual cost of testing in the District to be half a million. Is that a good estimate?

Jan: I don't know. He's a pretty reliable person. I imagine he worked it all out.

Bob: Some teachers said the district-built tests are too easy. I got the impression that the Superintendent kept pushing the committee to write items that show what all the kids have learned. That is likely to make the tests look so easy that they will be discredited.

Jan: Most of the people writing test items continued to write difficult items that would be useful for diagnosis.

Bob: The Superintendent told us himself that he wants to use the test information to get more support from the community.

Neil: The teachers would like that too, of course. Who wouldn't? But I don't think the Objective-Referenced Tests are what we have to worry about. It is the California Test of Basic Skills that has to go.

Bob: Are you basing that on the testimony or on your position in the NEA?

Neil: If we were going to base our findings only on testimony, we could have used tape recorders and I wouldn't have had to interrupt my vacation. I think I have been influenced a lot by what the NEA panel found out in Bakersfield, but other organizations such as the NAESP are coming to the same conclusion. The testimony here too was pretty clear.

Bob: Don't forget that several mothers were delighted to have the Basic Skills results.

Jan: They said that for years the District did not tell them how their children were doing in basic skills, but now they know.

Neil: True. But why do they want to know? There is no educational validity to those tests. They purport to measure what achievement is, but they do not. They do not help the teachers or anybody else understand the kids or help them to decide what to teach them. And they attach a stigma to the kids with low scores.

Bob: Two stigma deviations to some.

Jan: You are bad, bad.

Bob: Well, I agree with Neil, and I think that we should look closely at the question of validity. But I think that we will come up just as skeptical about the teachers' tests as about the California.

Jan: I was watching you working on a matrix of some kind, Bob. Does it have to do with the tests we are using?

Bob: Yes. Here's what I have so far. It's not much, but I think I can make it into something when I get home.

Jan: Can we put something like that into the report?

Bob: Yes. I believe it would help focus attention on the validity question.

Neil: Do you feel the report is going to be weak if we do not do more than analyze the testimony?

Bob: Yes. I wish we had more information. We don't know how much time has actually been spent giving tests, let alone getting ready for them and interpreting them.

Jan: The test program information in the Director's report is pretty good, I think. Have you had a chance to read it?

Bob: No, I've only glanced at it. But I wish we had been able to gather independent information.

Jan: Well, how will a table about the validity of tests make up for any shortage of information?

Bob: It won't. But I think the issue of test validity is one of the important issues for us to develop. I sense an agreement among us that the usefulness of these tests is low, whether or not they take up good teaching time. I think that we can discuss that lack of usefulness better in terms of the validity of the resulting scores.

Neil: I have a feeling this is the wrong way to go. I hope that you will have a statement that clearly says there are other reasons why tests should or should not be used other than their validity.

Bob: Right.

Bob: Bob here, back on the line.

Neil: The operator said you were having trouble.

Bob: I thought *she* was the trouble....

Jan: We've been talking about the validity table. You sent us an expanded matrix. Apparently you still feel that it should go into our final report. How do you figure that it will add to the report?

Bob: I do not believe the District can properly consider the questions about "too much testing" without looking closely at the utility of the tests. We have heard testimony from teachers and parents. We know that people are assuming the same tests are useful for many different purposes. I think it is important to recognize that no test has been researched and found to be valid for all those purposes and that some widely used tests have not been researched and found valid for any important educational purpose.

Neil: Why not just talk about the tests the District is using rather than about tests in general?

Bob: Perhaps that would be better. That would require some work that I do not have time to do. As you see, I do feel prepared to speak about tests in general. But, more important, I think that the Board should not be led to believe that they could choose some other tests for which validity has been demonstrated for those different purposes.

Jan: It is clear that you feel the validity for individual scores is different from the validity of means.

Bob: That is one of the most important points. The Superintendent is interested more in the means. The parent and counselor are interested in the individual student score. The teacher, at least occasionally, is interested in the response to the individual item. The validity is different at each level, but -- even for the best tests -- the validity has been researched only for the individual student score.

Look at the table I sent you. For establishing where a child stands in his/her group, or in a national group, or in some future predicted group, the commercial tests have been doing a pretty good job -- for groups as well as for individuals. This validity only holds, of course, with regard to substantive content of that test or with substantive content that correlates highly with that of the test. If what the test covers is not highly important to educational achievement, then -- even though validated -- the test is not going to provide useful information to a person who is concerned about achievement. Often a test is not what it appears to be. Often it does not tell us how accomplished, how educated a child is becoming. It deals with preliminary matters, skills, and knowledges, rather than the essential behaviors of educated persons.

Furthermore, as the table shows, we do not have good evidence that these tests guide us as to what to do in the best interests of the learner. I have no doubt that some teachers and curriculum coordinators can get that kind of use out of them, but most do not.

As you see in the third and fourth columns, for tests that deal with highly specific learning skills or subject matters, we do not have evidence that they are valid for any educational purpose the teacher or District may have. They may be useful, they may not; we just do not have the evidence.

It seems to me that what the table says is that validity has not been established for most of the tests and most of the purposes a district testing program has. Nor have they been found to be invalid. In this District it is reasonable that tests should continue to be used if enough people find them informative, helpful, and consistent with verifications that occasionally are made. I do not think the testimony we have heard so far supports any of the three tests they are using.

Neil: I do not object to what you are saying about invalidity of the tests. What I do object to is that your table says that some tests are valid for purposes that may seem reasonable but which actually are educationally and socially indefensible. I thought you were going to prepare the table so that a board member or other reader would not easily make the mistake of concluding that the tests you call valid are "good" tests.

Bob: Let me read you my statement to go with the table. It's a little long, but Ma Bell needs the revenue.

One of the issues of concern to test makers and test users for many years has been the validity of tests. People experienced with educational tests realize that any one test has a different validity for different purposes. A test may be highly valid for some purposes, but for other purposes that same test may have low validity.

A test with high validity is one that obtains -- with a high degree of accuracy -- the very information the user wants to obtain. A user does not, of course, want just a test score; he wants a test score that indicates something. A test is used at different times to indicate different things. The validity of the test each time depends on what the user wants indicated.

Educational tests do not measure directly the skills or understandings of a child, nor the effectiveness of a curriculum. They are used to indicate these things by measuring what children answer to a small selection of questions. Only a small sample of the many relevant questions is asked.

And even the total sum of all possible questions does not directly indicate what it is that the user wants measured. Educational tests are always indirect measuring instruments. These tests will have low validity if they are inaccurate -- but they also will have low validity if they are measuring something that is not a good indicator of what the user wants measured.

For some uses, the validities of even the best tests have never been demonstrated. Some tests have been used millions of times without a check on the validity of the most expected usage. For example, the validity of "reading readiness" tests as a guide to beginning or postponing formal reading instruction has not been established. The diagnostic uses of most tests are not based on "demonstrated validity." In other words, the technical studies to show that instruction is more effective when based on the test information have not been done. For many other tests and testing purposes the validity of the test is only assumed. Test developers and researchers have not yet demonstrated their validity.

During the fifty years or so that we have had these tests, the users have been interested in a relatively few possible uses of them. Recently, particularly with the arrival of the "accountability movement," many additional uses of testing have been proposed. It has been implied that tests that have been shown valid for discriminating among students would naturally be valid for assessing the effectiveness of teachers, verifying the quality of textbooks, determining the accountability of a district, deciding on a district's need for specially trained remedial reading teachers, and for setting national educational policy. It is possible that the tests will be *useful for these purposes -- but at this time the claims for such testing have not been backed up with validation studies.*

The purpose of these statements is not to argue that we should do such validity studies, but that we should not assume that they have been done. The purpose is to urge users of tests to resist the temptation to suppose that the tests will obtain complex information for us that has not yet been obtained elsewhere.

The validity of a test for a particular use is demonstrated by showing (usually in a carefully supervised statistical study) that improved understandings or decisions are reached by using the test. When test scores are used in combination with other observations to reach understandings or decisions, the validity of the test would be shown by the increase of effectiveness as a result of "adding in" the test information.

It is not unreasonable for educators to use tests for which the validity has not yet been demonstrated. They of course should use them with greater caution.

A test may be useful to teachers or administrators even when it has not been validated statistically. We sometimes speak of a "clinical" or "experiential" validity. Most test specialists are critical of such non-statistical bases for decision-making, at least if statistical validation is a practical alternative. I know only a few professional educators sufficiently knowledgeable about the curriculum and about the tests that they can use the test scores to improve instruction either at the classroom level or for the District as a whole. Our studies show that this is not true of most teachers and administrators. And we do not have a good way of knowing "for which users, in which situations" a test can be said to have a clinical validity.

Most people who are not well acquainted with testing have too high an opinion of the validity of the tests. The testing literature is filled with cautionary statements. They warn of expecting too much from the tests. But many persons, including experienced educational officials, let their yearnings to have instruction fully measured obscure these cautions.

In an effort to summarize estimates of the confidence we might place in tests for obtaining different information, I have prepared the following table. The statements of validity for the different tests are based on my experience, reasoning, and reading of the professional literature. I have submitted these estimates to several colleagues who have indicated that -- with perhaps a slight difference of opinion in two or three cells -- they agree with my estimates.

Some other colleagues have given additional advice. They have said that I should not circulate this chart because it will be misused, used to justify the abusive uses of tests.

Just as an accurate handgun can be used for immoral purposes, so also can a valid test be used for immoral purposes. Either can be hurtful through negligence. An array of test scores can be used to deny equal opportunity, to grant undeserved privileges, or to disguise bigotry.

Ranking students, assigning them to fast or slow groups, or treating them differently in school on the basis of predicted future success are potentially immoral ways of handling students. The educational benefits for these common practices are more apparent than real, and the social costs are potentially high.

For the first, second, and seventh purpose listed on the chart several test types have been demonstrated by psychometrists to be valid. But the social consequences of these uses was in no way considered as part of the check on validity. Each educator and each citizen (as well as each psychometrist) should be questioning the morality of these discriminations.

What we can see from the chart is that the tests have been shown to be valid for what was once their principal jobs: e.g., to indicate the relative standing of youngsters, to grade *them, to admit them to special programs, and to predict the level of performance at a later time. For almost all other purposes of testing, these tests have not been validated statistically. Some of the tests are too new to have gained a demonstrated validity. For some purposes the uses are too diffuse or idiosyncratic to justify the investigation. But for whatever the reason, the validity for most assessment purposes has not been demonstrated.*

It surely is as much a mistake to expect too much from tests as it is to fail to accept what help they can be. Many professional persons can benefit from the stimulation tests give to thinking about how to improve the curriculum. Many can use tests to orient students to their work and get them to work harder. And sometimes educators can actually use them to measure what they want to measure. Use of tests by professional persons with a full realization of the ill effects of blind discrimination, working to improve the opportunities for learning, should be given encouragement.

In many places there is a call for using tests to indicate the accountability of the teacher or the school system. For this use no type of test has a demonstrated validity. Such use of tests seems clearly unwarranted at this time.

Jan: I didn't get all of it, but it sounds good to me. Still, I'm beginning to think the whole thing is too technical for our report.

Neil: I think it is dealing with a relevant problem, perhaps as simply as it can be, but that it encourages the use of bad tests and does not give any help to tests that may be all right.

Bob: Do you want us to say that teacher-made tests are valid?

Neil: I think you should say that tests are valid when teachers can show that teaching and learning is helped by them.

Bob: Would you require more evidence than the opinion of the teachers?

Neil: I think for important decisions I would require what you call clinical validity and for minor matters I would suggest only that a teacher consider the testing carefully.

Jan: Yes, I doubt if we need to consider the statistical definition of validity. Maybe that is the main thing wrong with the table. It sanctifies this thing you call "demonstrated validity."

Bob: I don't mean it to. I think that people expect that tests are not purchased or mandated unless they have a validity that has been statistically demonstrated--and currently that is not the case.

Jan: What bothers me about Neil's comment is that it sounds as if teachers have better judgment about test scores than the test specialists.

Bob: When we are talking about using test scores to make decisions about what is to happen in a classroom, teachers do have better judgment than test specialists.

Neil: I agree. It is pretty clear that the value of test scores is very much dependent on who is using them for what purpose.

Jan: Well, we seem to agree on that. But how to get it said in the report without steamrolling the reader--or giving him false encouragement? And, by the way, I'll get steamrolled if I don't hang up and go teach my class.

Bob: I suggest you take a look at my statement in writing and propose corrections or get ready to decide whether or not to include it in the report.

Jan: O.K. In the meanwhile, have a good look at the most recent revision of recommendations I mailed you.

Bob: Got here as early today as I could.

Jan: We'll be meeting with the Board for a presentation in just four hours. Most of the report is run off and ready to collate. We have two or three small sections still to write, but the next thing seems to be whether or not we have final agreement on the recommendations.

Bob: I think they are excellent, Jan. You've done a fine job putting them together.

Jan: Bob, I don't understand the reservation you had earlier about making any recommendations at all.

Bob: I was talking about the danger of evaluators intruding too far into the decision maker's responsibility. The evaluator should realize that there is much in each local situation that he still does not know even at the completion of his study. Barry MacDonald is so concerned about the evaluator taking unwarranted advantage of his position that he suggests "no recommendations." In this situation it is pretty clear that we are expected, even obligated, to include recommendations in the report.

Jan: These recommendations are pretty weak. I think they say indirectly that the teachers are right: testing is probably taking up too much class time, but it is up to the people here in the District to work out something better. They would be more help if they were more directive.

Bob: No, I disagree. I think they would be too much help if they were more directive. So much of what we outsiders do in the name of service and counsel announces to the educators that there is a greater wisdom elsewhere, in Lansing, in Ann Arbor, back in Washington. But it seldom is true. The wisdom for solving the testing problems is right here in this District. What you and I can do is to give those who will work to improve the local situation some stimulation and some legitimation. As new problems arise, they may be able to point to one of our recommendations and say "Look. You've violated this recommendation of the Review panel." Then the so-called expertise from the outside is being used to back up the judgment of the inside, which I think is as it should be.

Neil: It's the same in the field of law. If laws are too specific, society is at the mercy of the creative crook.

Jan: Is that part of your reasoning for including the demonstrated-validity matrix, Bob?

Bob: Yes. I see these estimates of validity not as laws, of course, but as information. Even though testing is one of the great technical accomplishments of the educational-research community, we do not know whether or not test information can be counted on to be helpful in many situations.

Neil: You've heard of the Law of the Hammer: If a child finds a hammer, everything needs hammering. Something like that.

Jan: How do you feel about the validity statement now, Neil?

Neil:More convinced than ever that it should not go into the report. Just as Bob said, one can point to those high-validity entries for the published tests and say, "The Review Panel said these tests should be used." If it is your desire to include the matrix, then I want a statement of disavowal to accompany it. But I prefer to leave out the whole thing.

Jan: What would you say in your statement of disavowal, Neil?

Neil: That at least half of the purposes in Bob's table are not legitimate purposes. Whether measurements are in some technical sense valid is not important if, in some essential sense, the purpose of the measurement are contrary to the purposes of education. I would question these purposes:

To indicate standing of individual students with reference to norm groups.

To predict future standing of the individual in other reference groups.

To measure gain or improvement in skill or knowledge since a previous measurement.

To indicate the standing of an entire group.

To measure gain or improvement for the group.

To evaluate teaching.

Bob: Do you mean to say that, if teachers or citizens want to know any of those things, they should be told that those are not legitimate requests for information?

Neil: So much harm has been done to individual children and to all of education by people pursuing those purposes. The responsible thing to do is to point out this. What you have done in your table, Bob, is to focus on a less important matter and indirectly to encourage the use of standardized tests to accomplish these illegitimate purposes.

Bob: You may be right. I encourage you to prepare your disavowal.

Bob: I'm beat.

Neil:Yeah. And this limousine won't help. Ouch.

Bob: You think it went all right?

Neil:You can't do much at a dinner meeting.

Bob: The Board and the Association officers seemed to enjoy meeting together. Someone said it was their first, but they made it look like an old bowler's reunion.

Neil: Well, seems they gave some serious thought to the report.

Bob: They didn't get close to the validity issue.

Neil: No.

Bob: Your idea that the validity statements should merely be attached as a Panel "working paper" made sense. If the formal report had been twice as long, things would have been worse.

Neil: Maybe.

Bob: You expect to get this report circulated pretty widely?

Neil: It's up to the people here.

Bob: I don't see why they'd repress it. Everyone came off looking pretty good.

Neil: Can't tell.

Bob: Yeah.

Neil: I never figured out why you pushed that validity stuff so hard. You know that it came out making standardized tests looking better than they are.

Bob: Well, the estimates were honest. I think that we had to say some good things about the tests or we might have been called "bleeding-heart pinko radicals." The tower is too sturdy to bring down by logic alone.

Neil: Hmmm. But the teachers' tests are not part of your tower. Why were you so hard on them?

Bob: Of course, the tests are part of the same tower. Most teachers keep from petrifying partly by paying little attention to their own tests.

Neil: Hmmm.

Bob: In your "cautionary" statement you said that measurements people should get busy and develop tough-minded alternatives to testing. Do you believe it is even a possibility to measure educational attainment?

Neil: I don't know.

Bob: I have my doubts.

Neil: Hmmm.

Bob: Somehow people have come to believe that an education is the sum of knowledges and skills and higher mental processes. When you call a person an educated person, of course, you note that he or she has those things; but what you really see is that they have a great power of extension from one complex body of information to another. The connections they make are of high quality, they break new ground, they find new metaphors, they apply... You asleep?

Neil: Hmmm.

Bob: Harry Broudy and Karl Polyani called it "tacit knowledge." I expect we could predict who will have a lot of tacit knowledge in a subject-matter field if we pick a field at random; but if we pick a field that each individual is quite experienced in, then we couldn't predict it. It wouldn't be all that highly correlated with general intelligence.

Neil: Back on the old "prediction" road. Teachers don't need predictions. They need good measurements of what a child can do.

Bob: But it is so relative. Many skills correlate pretty well so there you can measure indirectly. Real educational power, real achievement is not sufficiently correlated from field to field so you have to measure it in the particular field, or sub-field. The fields get awfully small, but small fields are extremely important to a person.

Neil: So you would have 10,000 tests of subject-matter knowledge? Aren't you headed for the fine-grain measurement of the criterion-reference people?

Bob: They don't deal with grain at all. They deal with bits and pieces, not strands or complexes. They presume that with only a few skills and knowledges you can educate people. I don't think so. I think to educate people you have to involve them over a childhood and a young adulthood in opportunities and obligations to think, to associate, to explore, to play games with ideas. There are a million tracks to an education.

Neil: But what people want is a simple indicator that a child is making progress on some good track.

Bob: A simple score for the world's most complex game. Why do people long so for the simple?

Neil: Each of us sees it a different way. Especially those complex, ambiguous things like an education.

Bob: So the criterion is elusive, something you can't define, something you can't nail down.

Neil:And something you can't make a demonstrated validity table about. Airline? Yes, both of us United.

The Recommendations

The panel offers the following recommendations with the expectation that these findings will facilitate further study, review and improvement of the District's instructional programs. We recommend that the staff continue to use student-performance information as only one means to guide and to improve the District's instructional program. Other means should include new instructional procedures, in-service education and appropriate materials and media.

1. We recommend that the staff give greater attention to the limitations of standardized testing, especially when it is being used for purposes for which the validity has not yet been determined.

2. We recommend that in evaluating student performance greater reliance be placed on the expertise and professional judgment of teachers, counselors, and other specialists, support-personnel and less reliance directly on tests and other standardized assessment instruments.

3. We recommend that the amount of classroom teaching time used for testing be reduced, except when the teachers find the testing directly contributing to instruction in ways that justify the time and effort spent.

4. We recommend that the mandatory obligations of teachers to prepare statements of objectives, formalized criteria, and assessment tests be diminished -- and that teachers, administrators and other staff jointly accept such obligations only when in their professional judgment they find

that such activities contribute to the maintenance of a high quality of instruction.

5. We recommend that the entire staff -- teachers, superintendent, administrators, counselors -- assume joint and increased responsibility to communicate effectively with the community generally, and with the parents individually about student progress and educational activities of the district.

6. We recommend that the entire school staff become more aware of the ways in which assessment information is misunderstood by parents and others, and that they resist crude procedures such as mailing out student test results and offering uninterpreted school averages for publication, and that they make it as easy as possible for parents and citizens to get relevant evaluation information interpreted by someone fully qualified to do so.

7. We recommend that, whenever a testing program is operating, that an extensive in-service program be provided for all staff involved in developing, implementing, and interpreting evaluation of student progress in whatever ways it is measured.

8. We recommend an annual review of all aspects of the District's testing program involving -- directly or indirectly -- all who are affected by it.

9. We commend the staff for its beginning work on the Objectives References Tests in communications skills, but we remind them that the cautions about testing expressed in the report, apply to their tests, too.

10. We recommend that if ORTs are being considered in additional subject areas, their usefulness should be weighed against the time and effort which increase with each subject area added.

11. We recommend that the decision to use the *California Test of Basic Skills* be reconsidered in terms of its total costs and actual benefits to the District.

12. We commend the staff for recognizing the disparity between state goals and district goals, and recommend that resistance to the State Assessment Program be continued as long as its costs are seen to be higher than its benefits. We commend the staff for a clear understanding that achievement goals for each individual child are quite imperfectly indicated by District goal statements and recommend a continued higher priority orientation to the individual needs of each child.

13. We recommend that the total staff, and particularly teachers and counselors who have major responsibility for using the results of the testing program, be involved from the beginning in any further efforts to determine the purposes for which testing will be used, implementation, interpretation, and use of scores and will have a major voice in decisions to evaluate and revise the District testing program.

1976

The meetings of the Florida Educational Research Association had just ended. Jack, Bernadine, and I sat in the warm sun, trying to think of a title or metaphor. We weren't having much luck.

Yearning Power

What I wanted was a name for instant recognition of a certain kind of bureaucrat or what he does. The one who is forever reorganizing his office or appointing a committee instead of taking action. "Dervish" wasn't what I wanted.

I was a bit irritated. Jack had persuaded me that in my paper I had used the term "dilettante" wrong as if it meant over-specialized rather than amateurish. I had been bothered by several papers on evaluation matrices. My mind seemed to be on too many wrong tracks.

I wanted a word that pointed out that the official was able to avoid responsibility by getting people fixated on details, such as with an organization chart or task-analysis matrix.

The education people in Florida were apparently under a lot of pressure to think about curriculum and instruction in terms of specific objectives. At each school they found it necessary to make statements of what they were trying to teach and what student learnings should be tested. I had heard a couple of presentations about how the whole thing could be set up and monitored by computer matrices. That is, if you list all the students down this side and put all the basic objectives across the top, you can just check off what has been done and see what remains to be done.

Jack brought me back, suggesting maybe I was talking about a "butter and eggs man." Apparently that term had become popular in maybe the thirties to represent the guy who has his finger in every pie. Well, I rather liked the term; but it didn't do. I wanted to indicate clearly that the people were delayed getting the butter and eggs or whatever it was they needed.

Bernadine thought maybe I was thinking of a flim-flam man. But no, the bureaucrat I had in mind was not a confidence man. I was thinking of people who really believed that they were doing the right thing by rearranging their organization and planning new responsibilities. At least some of them. I wondered if there had been such a character in "How to Succeed in Business Without Really Trying." Jack said that there was the guy who gave the board members opportunity to watch the girl in the bikini while he was giving his spiel, but I was thinking more of a puffer fish than a shark. It wasn't my day.

It seemed awful to me to be representing the education of all the children in Florida in terms of a couple hundred objectives. Of course, I know that many educators and parents wanted a clear-cut approach and one could argue that you had to do the basics first if you wanted to get anywhere with the finer things. But it seemed so obvious that you could perpetually deny children the opportunities of learning the finer things this way. You give them shortcuts around complexity. Wasn't it important for their life to be intellectually enriched those thousands of hours in the school room?

They were really trying to do their job. They had promises to live up to. Like politicians. Nobody expects a politician to mean it when he promises things. Well, maybe he means it; but he knows that he won’t be able to deliver all he promises. But people listen to the promises anyway. I think they want to believe.

The school official can expect to be in office perhaps only three years now. Then move to another office, maybe to another place. Each move into a new responsibility he sees similar pressures. Still the problems are great. His predecessor had arranged things in a way that denied some problems their share of attention. So he reorganizes. A new action plan. New promises, new promises in the form of charts of objectives, organization charts, evaluation schedule, new this and that.

In technological societies, a design for a computerized data-system for educational program monitoring is especially attractive. The time-lag between the decision to have one and the test as to whether or not it works is a matter of, say, three years. The growth in design possibilities makes it easy to disregard previous failures. A new design can add legitimacy to the procrastination

People yearn for something better from their officials. They want to believe this time they will get their money's worth. They don't want a Don Quixote, tilting imaginary windmills and dreaming impossible dreams. But they too dream the impossible dream that the next reorganization will get work better than the last.

And the smart Dean or Commissioner or Superintendent, whether conscious of it or not, uses that yearning as protection, calling the team to draw up plans, showing a new list of expectations.

But I just couldn't think of a metaphor for it.

1976

I circulated this blurb to the CIRCE team studying public school science education in the U.S. for the National Science Foundation, a team including Jack Easley, Terry Denny, Rob Walker, Wayne Welch, Jacquie Burnett, Lou Smith, Mary Lee Smith, Rudy Serrano and Gordon Hoke.

On Seeing and Measuring

It is natural to see. It is natural to measure. Seeing and measuring are not the same.

But they are even more different than we suppose. The common notion is that when one measures one sees the same thing but sees its amounts. As if one were seeing through glasses having graduated-scale markings on them. Measurement glasses, however, do much more than scale the view. Much more difference there is between seeing and measuring.

There is a transformation from experiential perception to representational perception. The observer switches from actor to director. He/she gives up the direct impression of the thing, perceiving it no longer as another being, a whole object, a member of the physical populace, and perceives it then as a bearer of properties, or even merely as an array of characteristics. This is no small transformation.

When I find myself in the company of a rose I see. I do not see its redness, nor the Washington Monument its tallness, nor Professor Easley his intelligence. In order to talk about them -- and perhaps even to think about them -- I am always putting on the measurement glasses, and of course I see then, at least partly, each as a collection of properties: his brilliance, its height, its redness.

Getting ready to measure may be more like changing mindsets than putting on glasses. Taking vitamins, going on a diet, downing a coffee, or submitting to sodium pentothol may be more the analogue. They change mindset, changing one's ability to respond, changing one's experience itself. Now one fits into different clothes, into different roles, into different valuings. And these changes bring changes in strength and power.

The way most of my researching colleagues want to see the world is through the properties of things. The way most of my teaching colleagues want to see it is to see things as things.

Putting on glasses that focus on properties, scales, and amounts changes the perception. Perhaps only a little, as sunglasses do; perhaps a lot, as reversal prisms do. Whether the distortion is slight or great, whether the change results in more or less comprehensibility, the impression gained is different from that for the unaided eye.

I do not know whether the unaided eye is *more* or *less* likely to see truth. But it is important for me to realize that the perception of things with an orientation to properties, with an orientation to measurement, is "corrected" vision. Measurement is common and natural, but it is "corrected vision."

Whether or not such vision moves closer to truth is a matter to worry about. Many of us have not been worrying because we have been taught that when we measure. we are closer to truth than when we just see.

The difference between seeing and measuring seems small when Experience is the heat of the day and Measurement is the column of mercury in the thermometer. It is because of the commonness of looking at the thermometer, or hearing its amounts, and realizing the correspondence to our feeling.

For most of our measurements in education we do not have such a correspondence. Measurement it not just holding a ruler to what we see, but seeing something to hold a ruler to.

1976

By custom, Harold Gullickson was my academic father, and Ledyard Tucker gave me the strongest pull through my dissertation, but even though he caught me at 33 and left me at 49, Tom was my mentor for life.

Tom Hastings, Mentor

It is I, Bob Stake, wishing to honor before fellow American Evaluation Association members, my mentor, Tom Hastings, one of the pioneers of the evaluation profession. In the late 40s, Tom was a student of Ralph Tyler at the University of Chicago, with his fellow students, Lee Cronbach and Ben Bloom. Tyler was a specialist in Curriculum, but he and his four students moved quickly into the new field of student testing and on into the even newer field of program evaluation. Tyler supported the teaching use of behavioral objectives and was thought to originate goal-based evaluation, but he, Cronbach, and Hastings spoke vigorously for a broad base for seeking the merit of learning, teaching, and schooling. In writing "The Whys of the Outcomes," Hastings[1] held the roots of evaluation fast to comprehensive educational research.

At the University of Illinois in 1963, Hastings and Cronbach were joined by Jack Easley to create CIRCE, the Center for Instructional Research and Curriculum Evaluation. It housed the Illinois Statewide Testing Program until 1969 when they realized that the shift of testing away from student counseling

[1] Hastings, J. T., 1966. Curriculum evaluation: The whys of the outcomes. *Journal of Educational Measurement, 3*, 27-32.

to an accountability purpose was not compatible with the foundational purpose.

University Examiner Hastings served as assessment consultant to many campus, regional and federal projects, particularly the American Association of Geographers. He brought David Krathwohl, Phillip Runkel, Gene Glass, Ernest House, Douglas Sjogren, James Wardrop, Terry Denny, Gordon Hoke, and myself and many talented graduate students into CIRCE, and all of them, in turn. brought local and world groups together for discussion of testing problems and evaluation designs. And Tom was often the first to see a draft of Cronbach's writings and to nudge it more clearly toward the distinction it would ultimately receive.

A kind of one.

1977

From time to time, I was asked by funding or supervising agencies to review documents for merit, compliance, problems and the like. On this occasion I was hired by the highly-respected Carnegie Corporation to make a brief examination of a draft of an overview of instructional development. I don't remember who prepared the draft or the purpose of the review. I look back at it as revealing my growing concern about educational philosophy, about which I knew little.

External Review of the TORQUE Materials

10 September, 1977
Mr. Frederic A. Mosher, Program Officer
Carnegie Corporation of New York,
437 Madison Ave., New York, NY 10022

Dear Fritz,

I have spent a half-day on the TORQUE materials and have this to say: Intellectual support for this project is dependent, of course, on acceptance of a certain rationale for educational programming. Here it might be called a task-analytic, mastery, micro-structure or hierarchical rationale for classroom instruction. According to the rationale, there are certain lessons or tasks which when mastered and retained provide key aid in dealing with large numbers of subsequent lessons or non-school problems.

Many specialists in instructional technology, developers of criterion reference tests, and research people believe in this rationale. Most educators find it reasonable -- even though many of them do not rely on it for their own work. Some find it running counter to their humanistic leanings. I do not find this rationale supported by good logic, research findings, or professional experience. I do not believe the TORQUE staff has taken a good enough look at what education is.

One problem is similar to the problem of identifying what are the basic elements or building blocks of physical matter. Each time a theorist proposes that the atom, or the neutron, or the quark is the smallest possible element of matter someone comes along and shows it can be split into sub-elements. Dividing a task into its components seems also to lead one to an infinite regress. Few people agree that "those at any step” are in fact the building blocks for all sorts of school learnings.

A more important problem is that the list of key tasks that any comprehensive inventory identifies is greater than the total lessons available during the school experience of a youngster. So that any task, such as the measurement of length with a ruler, is not recognizable as a key ingredient in a sufficient proportion of school-learnings or life-needs to justify mastery of the task. There are too many equally unique and important tasks to justify a curriculum based on mastery of such tasks.

The TORQUE project aims at assessing the performance of students on such tasks. This measurement information is worth getting only if the task is in fact a key task, and if the measurement will help identify appropriate diagnostic activities, e.g., the proper time to go on to another lesson, and if the distraction of children from learning experiences is small. In my opinion, the TORQUE staff is making the claim that since measurement is very important and since measuring length with a ruler is a typical and simple example of measuring, then it is a key task. I think that it is quite desirable to give children guided experiences in measuring length with a ruler, but I do not feel that it should be treated as largely representative of measuring skills, and thus justifying instruction to a mastery criterion. And

I do not feel that it justifies an assessment of task performance as if it were part of every child's essential learnings.

Thus I am saying that though I fully agree that the general area of measurement is an area of important learning for all children, I do not believe that mastery of the task of measuring length with a ruler is important enough to justify the instructional time and assessment presumed by the TORQUE rationale. More measurement tasks should be taken up, I think, in the time devoted to the single task and its assessment.

There is a real question as to how much the teacher or curriculum coordinator can profit from assessment data. One school of thought holds that these practitioners, if properly trained and experienced, will be able to take assessment data and remedy learning difficulties and guide group learning substantially better than they could without the measurements. The other school of thought holds that it has yet to be shown that any practitioner can use assessment information in a diagnostic way such that the individual or the group ends up with a better education, and that the key responsibility of the practitioner is to give all the youngsters a chance to get at least a little bit into the problem, to have a bit of success with it, but to move on to additional tasks -- so that there is more contact with different tasks, content and problems. Here the expectation is, I believe, that it is important for the child to recognize the kind of intellectual demands there are, to meet some of them, but to be in a good position to relearn the skills later or to recognize what sort of special worker is needed to help with that kind of problem. Obviously, I suppose, I am in league with the latter.

I am interested in research that would support or discredit the assumptions of the former. I do not believe that at the present time we are training diagnosticians (particularly of a school or district orientation which I presume to require different understandings than diagnosis of individual student learning problems), even though with most accountability schemes there is a presumption that such diagnosticians exist. This question is not being examined within the TORQUE project, and of course there is no good reason why it should be.

If the rationale of the project is acceptable, I believe that what they are doing is quite reasonable. They are validating assessment against performance -- as they say, that is unusual and it is essential. I agree.

They are not, I would say, systematically exploring the limits of generality of their assessments. So we bump into such findings as the validity depends on whether or not you are using a ruler with a zero mark at the end of the ruler. The conditions under which validity exists may be very particular -- as I have suggested in earlier paragraphs. It seems to me that the staff is not sufficiently willing to explore these limits and possibly draw back on some of their claims as to the widespread utility of such assessments.

The actual techniques of recording validity in TORQUE are crude, which is okay, but unmindful of all the immense work that has been done on scaling and performance testing. It is embarrassing to read their claims as to how unique they are, and painful for me to observe so much defensive writing. They point to how inadequate other educational measurement is without acknowledging the simplistic notions of this assessment effort. That really probably is only cosmetic, and has little bearing on whether or not the project actually is a contribution.

At its present state and cost I feel the project is a disappointment. It is, in my judgment, based on an impractical rationale. It is not identifying a validation procedure that is beyond those used in performance testing nor is it showing how we could be utilizing assessment information in improving mathematics instruction. At least, I do not see the merit in the project that its staff is claiming.

Sincerely,

Robert E. Stake, Director of CIRCE

1978

Our largest, longest CIRCE project was Case Studies of Science Education*, funded by the National Science Foundation, at bottom, to assure contentious Congressmen that NSF was in touch with U.S. schools. We made ten-week visits to ten high schools and their feeder schools, across the country, for an ethnographic collection. This paper was presented at the annual meeting of the American Educational Research Association in Toronto, 1978.*

We prepared case studies of science teaching in U.S. elementary and secondary schools. Working with Jack Easley and me were: Beth Dawson, Terry Denny, Wayne Welch, Jacquie Hill, Mary Lee Smith, Lou Smith, Buddy Peshkin. Rob Walker, Rudy Serrano, Gordon Hoke and quite a few others. The final report was 20 times the size of this book.

The CSSE report was combined with final reports of two other national studies of school science teaching and discussed face-to-face with representatives of ten professional science education associations. One reviewer reported they spent 100 times extra with the CSSE case studies.

A Case-Studies-in-Science-Education Footnote

Case Studies in Science Education was a collection of field observations of science teaching and learning in American public schools during the school year 1976-1977. The study was undertaken to provide the National Science Foundation with a portrayal of current conditions in K-12 science classrooms to help make the Foundation's programs of support for science education consistent with national needs. It was organized by a team of educational researchers at the University of Illinois. Eleven high schools and their feeder schools were selected to provide a diverse and balanced group of sites: rural and urban; east, west, north and south; racially diverse; economically well-off and impoverished; constructing schools and closing schools; innovative and traditional. They were finally selected so that a researcher with ample relevant field experience could be placed at each. To confirm findings of the ethnographic case studies and to add special information, a national stratified-random-sample of about 4000 teachers, principals, curriculum supervisors, superintendents, parents, and senior class students were surveyed. Survey questions were based on observations at the case study sites.

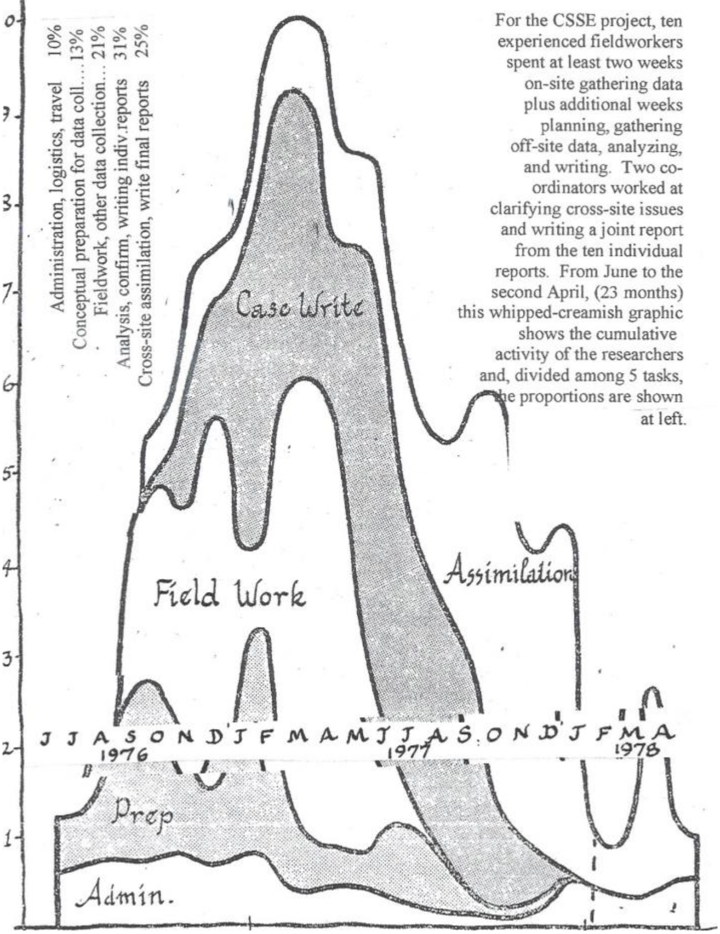
1978

The science-teaching multiple-case-study project, CSSE, described in the 1977 essay, ended in April of 1978. Toward the end, we worried some about whether our study had been rigorous enough. When we prepared our face-to-face presentation to NSF personnel we were pleased to learn that the head of the Education Directorate would attend. The presentation went well, over before we knew it. Now maybe a challenge to our efforts to measure the quality of science teaching across the nation by telling stories. When it came the Director's turn to speak, Richard Atkinson rose and said, "Well, let me tell you how it is in my daughter's school."

Different research projects use research time differently, but seldom is it reported or even calculated how a project uses its time. In this CSSE project, we did make the calculations and summed them graphically. Ours might have been a common way researchers spend their time, about a quarter for each of four tasks: Planning, Gathering Data, Analyzing Data, and Writing Reports. Perhaps a qualitative project, such as ours, with greater expectation of digging into unexpected findings rather than waiting to design the next projects, spreads the planning out more across the calendar. But for us, all four of the tasks occurred from beginning to end.

Use of Research Time in One Multiple-Case-Study

Figure 1978 - 1 - 1. Total Work Done by CSSE Research Staff: 121.3 Full-time equivalents during 23 months, 1976-1978



For the CSSE project, ten experienced fieldworkers spent at least two weeks on-site gathering data plus additional weeks planning, gathering off-site data, analyzing, and writing. Two co-ordinators worked at clarifying cross-site issues and writing a joint report from the ten individual reports. From June to the second April, (23 months) this whipped-creamish graphic shows the cumulative activity of the researchers and, divided among 5 tasks, the proportions are shown at left.

1978

A summer in Boulder, invited to teach with Gene Glass. We were sitting in Zucky's with Gene, some others from the University of Colorado. Gene identified Zucky's as a typical California restaurant. As I listened to people in the next booths, I realized everyone agreed that the governments, separately and collectively and trans-continentally are out of control.

Images of Governing

We had been in the state for more than a month. My friend Ernie -- he teaches at the University of Illinois (years later joining Gene at Colorado) -- had warned me about earthquakes, but no one had warned me about landslides. Yet here one was. Absolutely nobody believed that the governments were capable of governing. I had heard that New York City was ungovernable, but it finally was dawning on me that the whole country is ungovernable.

The next morning, I wrote it in my diary, underlining to make it special. The very first one, I remember, I called it Proposition 1 -- had to do with the images people create to make their world a cushy place to live, images like, "Anyone who really wants to work can find a job." Another was about schools and parents building up mathematics achievement to replace loss of the Church.

I wrote a new one on the morning of the 13th -- and to my amusement, it wasn't -- but a few days before, everyone around me was talking about Proposition 13.

Now, unlike Howard Jarvis, whoever he is, I didn't write that governments were malicious, or even negligent. Sometimes my father talks that way, but I've been around, and I know offices in Washington and Albany and Salem staffed with good, able people. Oh, they sometimes mess things up -- but person by person, and office by office, they try to do their job, and they try to make government work. I guess part of the reason I always thought of government as, potentially, another right arm, was that so many of my friends worked there.

If my friends and I had been born a generation earlier we would have run drug stores or shops of some kind. And if we had been born two generations earlier we would have settled on some of that land-bonus for Pennsylvania Volunteers. But in 1935 the land was dried up and blowing away, and the shops didn't have any customers. So, in due time, we went off to college to learn how, well, in a way, to live off the land, that is, from property taxes.

It was all right as long as we assistant commissioners and associate professors could believe that, in the long run, we were making things work or helping young people along. But the folks in Zucky's had no faith in us, nor in Jimmy Carter or Jerry Brown -- though one fellow said, "Maybe if a Franklin Roosevelt came around..."

I don't see how our kids can do without it. Industry and government together are providing a million new jobs a year. Private industry increases the number of order-takers at MacDonald's. Public government increases the number of Agency Field Representatives. As I see it, government makes some really nifty jobs. For many of our kids, government's still the employer of first resort: a law school, a community college, a reservation clinic, a labor and recovery ward.

MacDonald's puts out a better hamburger than Winnie’s mother used to, but costing much more than the nickel she charged. Neither better than Zucky's. None are the employer of last resort.

Well, I should have realized all this with Proposition 1. People invented the image of government, not to govern, not to establish protection, not to fight wars, not to insure domestic tranquility, but to make for some, the cushy place to work. The homesteaders had their day, the shopkeepers and we civil servants have had ours. Now I guess it's government’s turn.

1979

The politics of school reform has been an exercise at simplification. The problems are extremely complex, not amenable to the present "solutions" of the Governors and the President. A major change in values, management, budget and will seems necessary.

Education Unseen by Politicians

Different people have quite different ideas both as to what they want their schools to do and in what ways they want their children to become educated. Schools have a responsibility to offer parents choices.

However sincere the panelists, consensus on national educational goals is achieved largely by making goal statements: general, handsome, and free of pedagogical implication.

Most Americans are reasonably satisfied about what the nearby school is doing but are troubled about what is happening in the nation's schools.

The "state" has a vital stake in making its schools effective but it has never found a way of making them more effective by telling them what to change.

U.S. school achievement is as poor as international studies show it to be -- if that definition of achievement is used. Achievement test scores serve as a poor indicator of the job people see for the schools.

Achievement tests tell teachers very little that they don't already know about what their students have learned or about what they themselves might do to remediate or to teach better.

Achievement tests do indicate which schools and which countries have the best learners but they do not indicate which ones are doing the best job of educating.

In these times, the main value of standardized achievement tests is to inform (and pressure) teachers and students as to what learning is considered primary by the authorities.

The greatest reservoir of understanding as to how to make schools effective is the minds of experienced teachers.

Few teachers can verbalize most of what they are teaching and how they are teaching it. Nor need they. Most who do their jobs well do so without verbally describing the complexity of their work.

Teachers know they could do a better job but do not know how it could happen without restructuring their schools and communities.

One of the greatest obstacles to change in the schools is that many teachers only know one way to do their job and do not have the time, ability, nor interest to learn others.

The income from teaching is critical for maintaining the standard of living in more than a million American households.

People have too great a faith in the idea that if learners come to know the facts they will be able to use them; yet most of the use all of us make in our knowledge comes by informal experience rather than from formal teaching.

Curriculum experts are pretty much agreed that problem-solving, critical thinking and personal knowledge should replace much of the rote learning in our classes. Many teachers and parents do not agree, particularly when the free thinking invites criticism of authority.

In general, parents and teachers of children in the elementary schools deeply care more about kids developing personal and social responsibility than basic facts and skills and certainly more than about developing individualized intellectual skills.

The schools are no higher than third place in shaping what American youngsters know. Television and peer groups are more influential.

Primarily from the media, children have learned that to protest is good. The media have not taught them when protest is justified or how to use protest to get relief.

Most people approve of generosity to children having little opportunity but are belligerent when other children, even in affirmative action, are given greater opportunity than their own.

Extending opportunity by increasing access to classrooms and courses has regularly resulted in groups too heterogeneous to teach effectively and for a large majority of children, lessons either too boring or too difficult. Yet a key to reform is to honor, to defer to, the different cultures and experiences of our children.

Program evaluation techniques are presently too weak to tell teachers, parents and boards of education much about the influences of multicultural, racial, and impoverishment factors on teaching and learning.

A compassionate people, we have made our educational system greatly forgiving of failure, open to later opportunity. And so children understand that nothing needs to be learned today, that there's always another chance tomorrow.

Many schools are ineffective partly because a large portion of their students are not motivated to do what the teacher says. The teachers are unable to motivate or expel the unmotivated.

We frequently talk about helping youngsters be "all that they can be," to "rise to their full potential," but the fact is that there are no limits to potential, everyone can rise to a still higher understanding and proficiency. Learning goes on and on, as does forgetting. Attainment is a matter of will, their and ours, not of potential. This doesn't mean that anyone can learn anything in the time available.

The ambiance of the classroom varies immensely across the country. Curricula and pedagogy which work some places are certain not to work some other places. Heterogeneity can be a strength.

Teacher training institutions vary too, tuned largely to the schools and communities which hire their graduates. They don't do all their jobs well but collectively they provide a variety of teachers for the schools to choose among. Few Boards complain that the teachers are not being trained right.

The interest of young people nationally in becoming teachers is so low that soon parents of the average child in most classes will have a teacher less intellectually able than they. Implication for support of the public schools is large.

Innovative teaching ideas come not from research but from teachers with initiative. Research serves mainly to criticize (analytically and philosophically) and to expand thinking. It does not determine what is best.

The importance of good information as a basis for restructuring the schools is vastly overrated.

Leadership in schools today is defined more as a matter of negotiating power struggles of school and community than in interpreting what good education is.

Teacher unions are as positive a force as exists in the schools but seldom become visible other than when fighting school administrators for teacher benefits.

The philosophy of the American schools is one of local control but local school boards essentially follow superintendents, custom, and public outcry. Few boards reconsider what education is, thinking *that* to be a technical matter for the professionals. More control of schooling by teachers at the site is one of the few hopes for school reform.

1982

In the '80s, I had things to do in Washington, D.C. and it gave me chances to see the first of my to-be-eight grandchildren. Ozark Airlines, direct to Dulles, usually made it an easy run. Sometimes I flew into Baltimore.

On Being and Becoming

Last night I drove up Capitol Hill to see my grandson, Christopher, one year old, a being in his own right, a delight to the eye, pride's eye.

I can little evaluate what life is to him now. To me it looks like gobble, gabble and scoot. But I am impressed, as all always are, that he is different each day, and different in a progression of ways. He is indeed becoming. Of course, I do not know what he is becoming, but when he becomes it, I probably will not be surprised by the steps along the way.

Most of us in Education are so impressed by the developmental process that we impose it on our notion of education. Instead of thinking of school children as human beings we are more inclined to think of them as human becomings. Currently we are so frightened by the spectre of what some will become that we increasingly try to help all children become a little bit educated and a lot acculturated -- even at the expense of helping them have a self-gratifying, compassionate, lusty school life.

When approaching Capitol Hill, I road on Independence Avenue and leaving again for Baltimore, I road on Constitution Avenue. There are those two passways. Independence is the declaration that tomorrow is important, that we will live a different way than we have. The Constitution is the declaration of those things dear we would protect. A constitutionalist sees change as risky. We have much to lose. We should strive to "let it be." A developmentalist sees change as not only inevitable but wondrous and good, the vehicle of our emergence and remedy.

So thus the two, being and becoming, exist in one child, in one world, each part of the other, but in conflict. One cannot become without being, nor be without becoming. But each drains the other.

A parent problem is to balance custody and enticement. Even Grandparents affect whether safeties or new ventures predominate. We can try to constrain young people to the snapshots of the past, or applaud their groping beyond the world new to them toward worlds new to us.

Ours is a society that glorifies exploration, transition, becoming. A great deal of early 1981 attention was paid to the President's "Transition Team" as the new Administration came into office, and if you listen to the President's political rhetoric now, two years later, its ethic still is "transition." So much of past government was bad that all sacrifice should be made to put government and our people aright.

Ours is a school system that glorifies transition, preparation, becoming. Seldom is the question asked, "Are our children living well?" It is presumed they are living well if they are preparing themselves to live well some later day. When tomorrow comes it too is seen a time for further preparation or renewed remedy. Isn’t the balance wrong?

We are deluded by a notion that we can anticipate the ideals of tomorrow and shape our present to attain them. In spite of evidence to the contrary, we promise we will prepare our youngsters to cope with their future. As we suffer the drift of social conscience, the obstreperous economy and the miseducation of youth, we lower our aims for education, and for teacher responsibility but demand more compliance. As evaluators we are unwilling to accept human inability to shape social destiny. We are unwilling to accept our being. It is more than unbecoming, it is irresponsible.

Would that Christopher be spared the rod of excessive becoming!

1982

We were regularly under pressure to make our research objective and robust and to draw our students into allegiance. To my students less quantitatively inclined, I am sure I was not sufficiently open-minded, not as much I would be twenty-five years later. Sometimes we were amused by our own closed-mindedness.

Mail Survey Franking

Researchers continue to look for ways of increasing mail survey response. During a national survey originating at CIRCE in 1977, the research team felt it would increase returns to use return envelopes not only pre-addressed and "franked, but franked with an attractive postage stamp.” This is the report of the success of the effort.

One team member, Beth Dawson, felt that the best return would come if the stamp represented the personal concern of researchers, a humanistic interest in what the respondent had to say. The other group felt that the stamp should indicate the devotion of the researchers to the mission. The latter group chose the l3¢ stamp showing Lindberg's Spirit of St. Louis crossing the Atlantic. The first group chose the 13¢ butterfly stamps then available. Both were full-color, eye-catching issues. The butterfly pane included four species, allowing a further analysis of the results.

As a partial control a third stamp of the same denomination, that of an eagle bearing the American shield, was used with a small sample. Although full color and well engraved, it was then a more common stamp and not thought to signify the personal attention of the researchers.

A national sample of parents of high school seniors had been identified as part of Case Studies in Science Education, a project to be developed for the National Science Foundation. The purpose of the study was to assess the status of science teaching and learning in the nation's public schools.

The sample was obtained by drawing 35 high schools at random. Of these, 27 ultimately participated. A counselor was hired to meet with a "representative" class of seniors, each of whom addressed an envelope to one parent. The counselors sent about 750 questionnaires by mail. Of these, 401 were mailed directly back to CIRCE. The envelopes of 56 were somehow lost, leaving a sample of 345 for testing hypotheses about the influence of stamp choice. (This loss of data is the basis for the adjustments indicated in the table below.) Returns are shown in the table, assuming proportioned loss in the three designs.

	Approx. #	Adjusted-#	Number	Not	Percent
	Sent Out	Sent·Out	Returned	Returned	Returned
Butterfly	360	310	181	129	58%
Airplane	360	310	153	157	49%
Eagle.	30	26	11	15	42%
Totals	720	645	345	300	54%

The eagle control stamp was returned less than the experimental stamps, but the size of the sample was too small for this difference to be statistically significant (chi square = 0.93). The main comparison of stamps permitted by the volume of use indicated a significant difference (chi square = 4.8*) in favor of humanistic appeal represented by the butterflies. Orange butterflies had greater drawing power than yellow. Four species:

Orange tip (Anthocaris midea) o:	59 of 78 were returned
Checkerspot (EuphydrasPhaeton)o	49 of 78 were returned
Swallowtail (Papilio oregonius) y:	41 of 77 were returned
Dogface (Colias eurydice) y:	32 of 77 were returned

Chi Square = 8.7*

[Survey researchers are reminded to purchase supplies of state bird and flower panes (from Philatelic Sales, USPS,DC) while supplies last.]

1983

The current infatuation with accountability in Education is anchored in anti-person values. It is not the consumer protection it is billed to be, but rather an institutionalization of a system of subtle and not-so-subtle tyranny. Evaluators are the paladins of that power.

Anarchists Against Evaluation

The accountability movement in Education is unjust because...

... performance *standards* are typically arbitrarily selected;

...performance *measures* used are highly imperfect, indirect indicators of actual conditions;

... *measurables* are largely trivial and minor aspects of settings;

... *observables* are ambiguous to persons transient to the settings;

... *observations* are based on intrusions into settings which change the setting itself;

... *evaluators* are selected without check on bias;

... *values unexpressed* are often critical to the content of evaluation reports -- the hidden basis for the operationalization of the evaluation;

... *time* as a condition is often ignored by evaluators; and

... *autonomy* as a right of all parties studied is violated by *normative* review (evaluation).

Now therefore, we here assembled urge all like-minded colleagues to resist openly and covertly these dehumanizing efforts. Join us. Deny the claimed right to evaluate. Obfuscate real conditions. Attack persons, processes, and outcomes on all possible grounds. Demand representation of counter evidence. Resist with whatever tactics possible as individual circumstances warrant.

1984

A paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 23, 1984. It is based on a case study entitled, "An Illinois Pair," in a collection distributed by the Getty Center for Education in the Arts and the Rand Corporation.

Art Education and Critical Thinking

In recent years the American public-school curriculum has become more focused and simplistic. The appeal of back-to-the-basics, minimum competency testing, school accountability, equal opportunity, coupled with opposition to child-centered and life adjustment programs has restricted the range of classroom teaching and learning. Now, from many quarters, we see efforts to move emphasis to problem-solving, higher mental processes and concept-oriented curricula.

An unexpected ally -- at least unexpected to most researchers -- is the Getty Center for Education in the Arts. The Getty has worked to upgrade the art taught in schools. Some arts educators working with them -- Harry Broudy in particular, Elliot Eisner, Duane Greer, Brent Wilson, and others -- have stressed the importance of critical thinking and metaphoric imaging as part of general cognitive development. Using case-study research for The Getty this past year, we looked for practitioner interpretation of this new configuration of educational goals.

In Decatur, Illinois, we visited Centennial School. Kindergarten teacher Cole Williams directed attention to Horace Pippan’s *Victorian Interior*. "Is it balanced or unbalanced?" he asked. "Balanced," several children responded. "What if I cover up this chair?" "Then it's unbalanced," came the reply. "How would it feel, rough or smooth?" he continued. Later, he asked, "Who would live here? Kids? Pets?" "No, grandmas and grandpas," the kids answered back.

Staff development of art instruction in Centennial School has been provided by Project HEART, a service funded by the State of Illinois. Elementary teachers have been encouraged by central district administrators to volunteer for Nancy Roucher's in-service courses. Limited funding and the low priority for art education have kept this staff development from being mandated, even though state guidelines and districts goals indicated need for aesthetic perspectives and critical thinking.

Cognition

Although these teachers occasionally hear of Harry Broudy or see projections of his words, not many think of themselves as his following. They are tuned-in to a teacher facilitator, Nancy Roucher or Michele Olsen, and work with materials and activity plans provided by them.

From both workshops and their own practical experience, these teachers are aware that perceptual scanning is effectively taught with polar pairs. For example, is the painting balanced or unbalanced? Seldom apparent are the ultimate advanced lessons for interpreting and developing aesthetic expression -- partly because Broudy does not urge direct instruction of that advanced thinking. The ultimate goals press little upon consciousness -- the task for the moment is to develop a small array of conceptual skills, and to attain a "cherishing" of the parts of the arts thus made possible.

These teachers are aware that Project HEART and Centennial School, and their District as a whole, have been identified by the Getty Trust and are being watched. A few of them have read bits about the interest the Getty people have in restoring or vitalizing art education all across the country.

Almost none of them have given thought to the Getty call for curricular parity among art history, art criticism, and studio production. To them, this would be a fine-tuning adjustment. For the present the problem seems to be to get more modest accomplishments.

History is thought here to be a special subject, rather than a point-of-view about all subjects, not an essential component in all teaching. Yes, it may be useful, especially to stir some interest, to tell a story which has historical reference, but an emphasis on history in teaching art is infrequent. This is as true for exemplary teachers as others.

Occasionally there is attention to local history, such as in Gary Olsen's sessions on architecture. When history does appear in the lesson it usually pertains to the history of a work or the life of the artist. Less often is one pointed to the history of art -- its periods, its schools -- its fashions, its concepts, its paradigm shifts -- or to relationships between art and social, scientific or other histories. It would be unusual for a Project HEART teacher to try to illustrate the problem of an artist, e.g, Monet, struggling to come up with a new medium or technique for an expressive quality he wanted. That very special interaction between critical judgment and personal expression is an essential relationship in education generally, one sometimes well taught through art education. But it seems too complicated for most trainers of teachers to figure out. Not surprisingly, it appears to remain beyond the scope of teaching in Decatur. In one of our other magnet schools the separation of aesthetic education from studio production has had the effect of keeping critical thinking from being overwhelmed by the vitality of production, but also of isolating criticism from production.

Criticism appears not in the mien of connoisseurship, as Eisner would have it, but more as analysis, or rather as preparation for analysis. Of course, not all criticism and critics are analytic, but identification of properties (by the teacher and student) is considered to be a first step in understanding and communication. Holistic or metaphoric criticism is encouraged to a lesser extent. What all the children in these classrooms become accustomed to is the description of an art object in terms of its properties. Perhaps later, movement toward parity of criticism and studio production.

What a child likes, is more or less treated as unimportant. Personal preference is of course encouraged in production, but not in criticism. A major issue remains to be faced when the question of "whose" standards -- history's, the teacher's or the student's -- comes up. For the time being, perceptual scanning is taught as objective, not subjective. Project HEART teachers are pleased to be teaching something substantive in art class.

The case study prepared by Robin McTaggart and myself portrays educators concerned about art education in Champaign and Decatur, Illinois. They are not uncertain about how large is the task, nor how small the opportunity. Developing a conceptual approach to school art is not widely under way. The "basics" remain dominant. When these teachers hear about student recognition of expressive qualities in art and about enriching the images of all thinking, for all students, they are hearing about achievements well down the road. Yet, with Getty help, the Project Heart teachers have raised their aims. They have been sobered by the District's inability to provide substantial impetus, disappointed by the lack of interest in critical thinking of most classroom teachers. But as most teachers do, they largely ignore the gap between real and ideal, and work when and where they can, to lay a certain groundwork for critical thinking.

1984

I had a three-month scholarship sponsored by the Fulbright Foundation at Brazil's Universidad Federal de Espiritu Santo, participating in its Disparities Research Project. This case-study report describes an off-campus rural school trip accompanying Professors Beth Gama and Gilson Pinciara Sarmento.

A Brazilian Field Trip

Anchieta County lies between the coast and the mountains; the south-flowing Rio Enchante divides it east and west. Fishermen pull into the river to dock their single-horsepower two-person boats. A fish market with 4 or 5 shops is nearby.

After a mile or so of Anchieta's narrow urban orientation, the county seat's main street at each end becomes a paved road, one headed to Rio, the other to the state capital, Vitoria. The other roads running through the foothills and into the mountains are dirt roads but, until the steepest, well maintained. Car traffic is light, trucks occasional. Power lines crisscross the county and many houses show a TV antenna aimed at Rede Globo's booster station, bringing in the nation's major network.

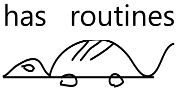
Houses are almost entirely of tile construction on a concrete slab, cream or tan in color, with an ever-blackening red tile roof. Some farm places have outbuildings but not most. One-room schools and chapels pop up every couple of miles. The foothills are increasingly stripped of forest and bush, exposing pockmarks, old anthills. Cattle, particularly Brahmin, appear to overgraze the hillsides. Coconuts, bamboo, sugarcane grow in the draws, papaya, bananas and coffee on higher ground. A few chickens are seen, dogs more common than cats, horses wait patiently for their rider.

The ground is badly eroded where roads have been cut through, exposing an earth bright orange, red or light tan. Granite outcroppings are common. As the mountains approach, the roads narrow and roughen, but the houses get a bit fancier. One never seems far from a place to buy popsicles or beer. Roadside advertising is infrequent. Main highway signs caution drivers to protect life.

At the public schools (all but one in the county, Grades 1- 4), the visible signs of curriculum and pedagogy are few. There are teachers and workbooks, desks or chairs and a blackboard with something on it. There are no textbooks, bulletin boards, artwork, children's exhibits. One room has two religious pictures, framed, probably placed there when the building was built. Almost the only wall charts are teacher-made, the most common are phonetics with drawings: ta

One teacher has cut out magazine pictures, making several decorative posters. Only one room has a flag, two flags, in fact, paper, the bright green national flag. A local pioneer, Duque do Caxius, appears framed in one room, unframed in another.

Some rooms have been cleaned today, by the teacher with student help. (The 10-room town school has two women custodians.) Other rooms appear to have gone many days without. Orange peelings on the floor and trash out the window are not uncommon. The only fresh poster (appearing in 3 schools out of 8) is a cartoon character, arms around a school building, admonishing readers to "Love your school," meaning in part not to vandalize it. (The town school has been badly vandalized over time, and not repaired.)

Whatever the problems of these schools may be, the personnel resource of them seems a major strength. The teacher has routines for the kids which indicate a considerable  involvement in arithmetic and language (long division and syllabication and phonics most evident). Workbooks are standard fare, in most schools each student's book bag carries several. Group work is evident in one-sixth grade class, though in another, students seem unfamiliar with the routine of a group assignment just made.

The teachers are not puzzled by what they are supposed to do or what the children are supposed to do. They are frustrated by the obstacles, mostly the shortage of materials. A good relationship, teachers with students, and within those two groups as well, is evident at each site. Portentiousness is not the teacher's demeanor – readiness to tell us about the school's activity and need is.

Socialization in the classroom appears quite healthy (in contrast to upper grades in U.S. schools where teacher and students often have contentious relationships). Here (excluding Grades 5-8) the teacher and students have a common goal, and show respect for each other. The children work soon after being told, and have admirable attention-span to the task. Some peer assistance is apparent, but is not taken advantage of by any teacher we observed.

Taking tests is being routinized. On one blackboard a second grade "teste" consists of four items in cursive. (On some end-of-year tests, these children must get 80% right to pass to the next grade.) In this county, the educational coordinator gets teachers together to make up the test, then sells it to the pupils (15c) each to make money for office operations. The tests are mimeographed with the cover sporting a cartoon character, hand-colored neatly. This office also sells maps of the county, ours showing the correct location of some schools, but not all.

The signs of school poverty are many: Barren walls; the smallest, cheapest tablets for pupil writing; teachers providing their own chalk; no water in the toilets except that brought by bucket (but always two closets); no stove to cook the government-supplied hot lunch (but pupils happily gathering wood for campfire cooking); desks over 50 years old of a style of 100 years old; no bulbs in light sockets in a school with light sockets (most are not electrified); only six new brooms this year for almost 50 schools; no paint since construction; children unable to buy the “required” emblematic T-shirt; kids walking miles to school, almost none with bicycles; the shared eraser, dime-sized; some using pencils an inch long; teachers charged for government-supplied grade books, which sometimes are sold to buy fuel for the cook stove (where there is one).

The signs of spirit are many: Children happy, teachers involved, most animated; blackboard announcement of today's date; children working quickly, comfortably when the teacher tells them to; children interested in each other, sometimes helping each other with lessons; eighth grade boys clustered about the gate so that the older girls have to squeeze through to leave; well-groomed teachers walking a mile to catch the bus home; children organizing group games, singing.

There are political signs too: More resources at some schools, partly (apparently) based on which teachers are closer to the county mayor, or whether or not the local landowner's children are in school; poorest school is a Black school, where only 6 of 20 attend today (the Black school down the road a mile isn't open at all now because they have no teacher), the children here don't turn around and look at the three visitors for ten minutes, the teacher said they thought we might be the police; when asked who is the smartest, who helps the teacher most, the spontaneous answer is, "Everybody's the same," school shirt ideology is to show them all equal -- the Minister of Education corrects school officials saying children must be admitted whether or not they have the shirt -- so again we see classes, "shirted" and not. The federal shipment of food to the state for distribution to county schools said to arrive just prior on an especially hot summer, was badly stored, spoiled. Some relate timing to the fact the opposition party is in power here.

Signs of isolation: Although these rural schools are within 20 miles of the county seat there is little communication. Some teachers say in the last two years they have not seen the coordinator in the field; schools have no phones; even advertisers don't look their way; still, the older kids know who Michael Jackson is.

Signs of ethnicity: Most schools display great gradation in complexion and hair color, from fair to very dark, but most kids are rich brown, with bright eyes a standard; one imagines that one sees Nordic, Mediterranean, Portuguese, African children; they show little ethnic cliquishness, some; some aggression is apparent in one school, among boys, with youngest Blacks appearing to take more than their share of hazing, the Nordics the most aggressive; bigger girls of both ends of the color continuum control them with sophisticated moves, small threats of exclusion.

Now I am sitting in a school across the street from the beach in Uda, a small fishing village. This is a second-grade room this afternoon, in the morning it is for some other grade. The regular teacher is away and a qualified substitute (in some places a member of the family shows up to teach) has been called in. This one is a young woman of perhaps 22. She says she likes to teach but seldom gets the opportunity.

Just now a child is reciting, reading from a workbook. (It may be that visitors get to see children performing standard roles more than their teachers.) The workbook asks for the recognition of syllables. Her reading is lively, steady. Now a boy reads. He is not so able. The class chatters a bit; the teacher shushes them (in Portuguese (?) it's more like "ssssst"). Several children are working at their notebooks. (Behind me through the open window several older children check out the two of us visitors.) The reader murmurs on, finishes, relieved. Now a third reader. The children have all been over this same page. (These are the highest quality workbooks we have seen.)

On the walls, bare in other schools, are posters, handmade in the coordinator's office, indicating vowels; something about the Duque de Caxias, Patrono do Exercito Brasileiro, etc.; several from magazine pictures. On the blackboard are chalked exercises, as at other schools, in the teacher's longhand. Here we have 20 kids, half boys, half girls, most of them chocolate brown, with all colors of hair from blonde to jet-black, kinky to straight. Their eyes are large and bright. They are curious. We are an interruption, not unwelcome. They impress us with their penmanship and language. A new exercise now goes onto the blackboard:

---greja ---relha ---vo ---viao, etc.

Now we are near the mountains, visiting the county's only public secondary school, Grades 5-8. There are 4 female and 1 male teacher, all in their 20s or 30s. Teachers and maybe 100 students finish their noodle soup and the cook collects bowls and spoons. The kids are clean, healthy-looking, well-clothed in jeans and T-shirts. There are no Blacks here.

Today the seventh graders have math, geography, Portuguese, science and physical education. Religion is taught once a week. We go to science class. A pert young miss reads a group report on blood circulation, drawn from books brought by the teacher. Only one boy seems to listen seriously, but the others remain subdued. They find the visitors only briefly worth scrutiny, but there's a sense of "things are different today." Two other reports follow, one on respiration. The group doing "human cells" hasn't finished theirs yet.

School is out now for some, but the seventh grade goes to P.E. First a run around the soccer field, then contests and games, including a version of "drop the handkerchief." Boys tend to choose boys, girls girls. Notice, the teacher objects to the youngsters helping run things. The youngsters enjoy themselves in spite of her seriousness -- perhaps a seriousness more for us outsiders. When asked why they like school they say, "It's good to be with the teachers and the other kids." Two boys, book bags swinging, head for home, a two-hour walk to the "furthest mountain," that one over there just now starting to hide the sun.

1986

I first told this story in a technical meta-evaluation federal report, then in my 1986 book, Quieting Reform.*[1]* Quieting Reform *might have been my best evaluation report, even though it tells of a failure I was a partly responsible for.*

I think most of us professional program evaluators of Charles Murray's time and now, still, 36 years later, have neither the mindset nor the resources: to go beyond simple indicators, to acknowledge Tolstoy's multiple causes, and to do justice to the range of stakeholder concerns.

Failure to Evaluate

Seldom hangs a failure alone. When we fail to evaluate a program well, the causes will probably be several. It is reasonable to expect some shortfall by the evaluator, of course, but also from collaborators and practitioners. And administrators and stakeholders, and close friends and tormentors.

My story of evaluation failure involved some pretty fast company, starting with Jimmy Carter. It began when he was governor of Georgia and came with him to Washington. His favorite program for government delivery of social services was Cities in Schools (CIS). He had almost nothing to do with the evaluation of CIS but -- in failing to protect it from prevailing notions of accountability -- he, alas, was part of the failure.

The evaluation also was undercut by Bill Milliken, the national Cities in Schools director, the leading organizer in the initial cities, primarily Atlanta. Bill was a charismatic social advocate, something of a sectarian Jesse Jackson, sensitive to

[1] Stake, R. E., 1986. *Quieting Reform.* University of Illinois Press.

the plight and disadvantages of urban youth, skilled in bringing young professionals together to help other young people. He thought it the work of evaluation to pass judgment on instances of visible merit, not just to find correlates of it in large-group sociometrics. He did not object to using the easy marks such as class attendance, grade point average, and avoidance of trouble with the police as indicators of rise to social responsibility. Not objecting can be a contribution to failure.

The failure occurred under the operational direction of Charles Murray, a brilliant social scientist with radically conservative tastes and driving personal aspirations, a chief researcher at the time at the American Institutes for Research (AIR), which held the federal contract for the evaluation study. AIR had been involved in studies in Southeast Asia near where Murray had been in the Peace Corps. Political ideology has not been a predictor of technological or social service failure but Murray's investment in scientific management such as Robert MacNamara's Planning, Programming, and Budgeting System (See Enthoven, and Smith[2]) had little to recommend it for analysis of social change.

AIR President Paul Schwartz urged an evaluation design needing an "incremental ethic," with data collected on gradual changes away from unwanted behavior and added to data needed for the usual correlation studies of treatment and outcome variables.[3] This elevated the role of evaluator as social scientist and therapist while diminishing attention to stakeholders and the quality of what the program people were actually doing. Such concentration on student traits and incrementalism, I contend, contributed to the gross shortfall about to be described. Murray and his advisors settled for measuring initial increments only, and failed at that.

[2] Enthoven, A. C. & Smith, C. K., 2005. How Much Is Enough? Shaping the Defense Program, 1961-1969. National Book Network.

[3] Stake, R. E., 1986. *Quieting Reform.* University of Illinois Press.

Cities in Schools was a collection of urban "youthwork" projects, first conceptualized by Milliken in street academies for dropout students in Atlanta. The goal was to find the most estranged youth and bring them through the schools into mainstream urban society.

Zealous staff members, most of them "out-stationed" from a rehab clinic or fire station or other agency of the city, worked intensively and personalistically with individual youth, and gradually extended to other members of family or gang. The meeting place increasingly would be the school. In collaboration with curriculum coordinators, the youth worker would actually take on formal teaching responsibilities, but maintained the personal coaching in social responsibility, matched to youth and agency.

As designed, some social services would thus pass through the schools. The schools scooched over a bit to make room. Integration of social services and education became a second signature of Cities in Schools. It was promoted as opportunity to enhance the delivery of social support to families and isolated adults as well as to the dropouts.

When Charles Murray designed the evaluation, he recognized that delivery of city-wide social support through Cities in Schools would not happen if the dropouts were not effectively served. "Measure what is measurable." He organized his data collection around the quality of their participation as measurable at the schools. He prioritized (1) school attendance, (2) grade point averages, and (3) avoidance of trouble with the police.

By 1977, Cities in Schools had "gone national" with several sites federally funded. The National Institute of Education (part of the U. S. Office of Education) held responsibility for evaluating the federal effort. Its "Request for Proposals" called for a "stakeholder" evaluation, an advocacy of NIE evaluation monitor (and friend of mine) Norman Gold.

American Institutes for Research won the contract and assigned senior researcher Murray to take charge of it. It became Murray's last research contract before becoming a professional writer and publishing widely circulated political critiques, *Losing Ground*[4] and *The Bell Curve.*[5]

For the evaluation, NIE insisted on a prestigious advisory committee and contracted for it with the Evaluation Research Society, a Harvard based group with strong federal interests. In 1986, ERS was a forerunner of the American Evaluation Association, before long to be the world's leading evaluation professional society. I was named to the advisory committee but withdrew after a meeting or two when NIE contracted with me to provide a meta-evaluation (evaluation of the evaluation) of the Murray's investigation of Cities in Schools. Additionally, a large set of contracts was made by AIR with individual schools to be studied to provide student demographic and impact data.

So there were at least six parties needing to be accountable: NIE, Cities in Schools, the public schools hosting it, the AIR evaluators, the ERS advisors, and the meta-evaluator, each of whom had a potential role if the evaluation were to fail. Not only were contracts to be fulfilled by law, regulation, and ethics, but with findings that would be understood in terms of custom, colleagueship, and aspiration.

The advisory committee was vigorous in its press for high standards of evidence. It pushed Murray for hard data, satisfied that program merit would correlate with school grades and avoidance of trouble with the police. It approved the plan that studying CIS operations in a few schools (but no street academies) in three cities, Atlanta, New York City and Indianapolis, could represent CIS city-wide and nationally.

[4] Murray, C., 1984. *Losing Ground.* Basic Books.

[5] Murray, C., 1988. *The Bell Curve.* Basic Books.

A year went by. The youth workers continued doing "their thing," seldom leaving a paper trail. The schools were buffeted by a teachers' strike, confusing regulations, turf disputes, problems with untrained caseworkers, and management priorities unacceptable to various constituencies, but these were not extraordinary events for CIS and all of urban education. The youth-workers taught, coached, counseled, cajoled and maintained an *esprit de corps* at least on par with the nearby schools. A small few research observations were made of CIS activity.

The schools did not provide Murray with the data for which they had contracted. He acknowledged he had not been demanding enough as a data manager. Often enough, he urged his sources to suggest and supply other evidence of program success, regularly disappointing him.

At evaluation's end, in his AIR final report to NIE, Murray made two main points:

1. Cities in Schools impact on the targeted urban youth was imperceptible.

2. Federal efforts to sway the most wayward youth should await reconceptualization of the federal role in welfare.

The ERS expert advisors supported Murray, reasoning that had there been major service to youth done, it would have shown up even in the meager data collected.

It remained unknown if Cities in Schools actually failed as a social service. Success, if it existed, was not perceptible to the evaluators. The evaluation failed to comprehend what was happening. Where were the bulwarks protecting evaluation from failure?

CIS: It didn't assist the evaluation as much as promised, and didn't insist on evaluator portrayal of high-quality personalistic assistance to youth and integration of social services.

NIE: It monitored, cajoled, and despaired, but to little avail.

ERS: It endorsed irrelevant evaluation standards, so in that regard, ERS was a significant part of the failure.

AIR: It failed to search for evidence of high quality personalistic assistance to youth and integration of social services, settling for weak "approximates," so in that regard, it too was a major cause of the failure.

The meta-evaluator: I didn't comprehend the weakness of the evaluation until a year too late. I had talked with Charlie several times a month, little acknowledging the illogic of AIR's rationale. I also was to blame for the evaluation's failure.

Charles Murray sought program quality with faulty criteria, using just three, presuming them to be indicators of true programmatic success; criteria having low validity and faulty administrability. His basic contract authorized the simplistic criteria he used. Along the way, when he asked Cities in Schools to come up with additional indicators, Bill Milliken summarized for him what was happening at individual sites, but the CIS testimony remained too personalistic, subjective, and situational for AIR and ERS evaluators. Charles Murray needed to have taken Jimmy Carter seriously -- but perhaps Carter's focus was already on Ronald Reagan. Tolstoy reminded us that the fall of an apple can be seen in multiple ways, which some would call "causes." He asked*,*

> *"Why does an apple fall when it is ripe? Is it brought down by the force of gravity? Is it because its stalk withers? Because it is dried by the sun, because it grows too heavy, or the wind shakes it, or because the boy standing under the tree wants to eat it?"* [6]

[6] Tolstoy, L., 1869*,* 2007. *War and Peace.* Random House.

Failures too can be seen in multiple ways, with multiple contributions. Compacts of failure are to be expected. Short-sighted is the evaluator who attributes virtue or misadventure to only three indicators. Short-sighted is the profession that endorses simplistic explanations.

1987

And there was no pandemic to blame it on.

American Education, 1987

Education continues to be problematic in the United States. It serves much less well than we would like as an institution for the preservation of cultural values, for preparation of youth for adult life, and for exploring knowledge frontiers.

Considerable attention has been given the plight of public education. State legislatures have responded to reports of unskilled graduates in diverse ways, particularly in the creation of assessment systems. It has been presumed that were we to measure more effectively our shortcomings we would know how to remedy them. It also has been presumed that basic skills need to be well developed prior to introduction of important content, experience, and interpretive responsibility. Our University of Illinois studies show both to be questionable assumptions.

The curricula of the nation's schools have become increasingly focused on uniform statements of objectives. That is good and bad. We have reduced irrelevancy, but complexity as well. Were we able to voice effectively our deep and complex expectations of education; were we able to serve individuals well by concentrating on common objectives rather than on combinations of common and individually tailored objectives; goal statements might enhance the curriculum. With the objectives and syllabi we now compose, we are simplifying, narrowing, and moving away from what literary critic E. D. Hirsch[1] has described as "cultural literacy."

One of our American virtues, yet one of our problems, has been "egalitarianism." We have increased access to educational opportunity for minorities and the poor. In assuring access we have opened almost every class to any student, regardless of readiness and regardless of willingness to participate. In many schools rich and poor, student "work ethic" is problematic. We now have pupil heterogeneity and a rejection of instructional activity, both of which hobble teaching. Ways need to be found to guarantee each willing learner an environment free of a dominant student grouping unsympathetic to, even obstructing, the teacher's intentions.

The reform efforts of districts, states and the U. S. Department of Education have been intuitively and politically attractive, but counter-productive. Articulation (content patterning) is overrated. Assessment has led to "teaching for the test." Public education itself is at risk partly because corporate management principles (common to business and widely acceptable by school administrators) overstress routine and conformity, and undervalue an intellectual environment.[2] Instruction occurs inside school and outside school, thanks to parents and mass culture -- so gross illiteracy is avoided -- but students, teachers and the educational ideology of our people are intellectually impoverished. The competency of today's teachers is not sufficient, but greater than that of our school and central administrators. Perhaps the highest priority should be -- within broad limits and with professional assistance -- freeing teachers to teach what individually each can teach best. Grand efforts to reform the system are hurting more than helping.

[1] Hirsch, E. D., Jr., 1987. *Cultural Literacy: What Every American Needs to Know*. Boston: Houghton Mifflin.

[2] Feinberg, W., 1983. *Understanding education: Toward a reconstruction of educational inquiry*. Cambridge University Press.

1987

At the annual meeting of the American Educational Research Association I was a respondent to four papers on the topic, "The Act of Teaching: Theatrical Perspectives." My questionably timed response was as follows:

The Ionesco Clock

Just a week ago I was driving in Australia with Stephen Kemmis. He asked me if I understood a road sign, a rather new post marked "M130." I guessed it was the highway number. "No, it means we are 130 kilometers past Melbourne." Stephen lamented the highway department philosopher who insisted on telling us how far we had go -- rather than how far we had still to go.

You may think it unreasonable for anyone to assume *where* Stephen and I were headed, but no more so than to assume from whence we had come. It made just as much sense to assume we were bound for Adelaide as to assume we had come from Melbourne.

In America on roadways and in classrooms we assume we know where we are going. In modern theatre, Madeleine Grumet[1] reminded us in her paper this afternoon, we expect the unexpected.

As chair of this session I have kept time. As I sat here I was a clock telling how much time is left. I recalled that basketball has such a clock -- but education has not. Apparently there is little market for a clock that tell how much time is left? It's better that way/

[1] Grumet, M., 1976. *Toward a Poor Curriculum.* Kendall/Hunt.

I'm not speaking *against* time passed, against a sense of history. I like to muse about where we've been and how long it's taken, but when I get down to this session's topic, "Act of Teaching," I'd like to know how long we still have. *My* opportunity to say how long Education can last like this, seems to be slipping away. At these meetings I've heard said, "It's later than we think." Some have implied, "It's already over. We missed our chance." When we talk of changing teaching, I'd like to know if we have any time left.

Some say, "It depends! It depends on what we are going to do." I agree. It depends on what we are capable of doing. And the resources. It depends. But for much of what we have *promised* to do, there's *no* time left. For many goals we never did have the clockwork to do it.

I've looked and listened in these halls. Much of the time we seem headed nowhere. Would it be refreshing to know that we have forever to get there?

No, seriously and thankfully, we are headed many places at the same time. Our new clocks could show *that*. Just as bank clocks tell time and temperature, in Fahrenheit and Celsius, our many deadlines could be tolled. Our deadlines slide, but the better clocks would show that too.

A clock is not a clock is not a clock. Just as Grumet tells us -- the theatre is not the theatre is not the theatre. Just as Miriam Ben-Peretz and Sarah Scheinmann[2] tell us the classroom is not the classroom is not the classroom. It is all at once the real, the unreal, the surreal. In it, or watching it, we drift from *seeing* stereotype to deviance to absurd, from *being* stereotype to deviant to absurd -- from conversion to revolt.

[2] Ben-Peretz, M. & Schonmann, S., 2000. Behind Closed Doors: Teachers and the Role of the Teachers' Lounge. *Anthropology Education Quarterly.*

This afternoon Risa Whitson[3] drew our attention to dramatic connections among Clifford Geertz, Fred Erickson and Edmund Burke, and pointed out need for post-structuralist insights and dialogues among teacher psyches and student psyches. Paula Salvio[4] ticked off the spitball scene, embodied in text and teacher. Madeline portrayed Mr. Patton's scene where the teacher simultaneously played aspiration and denial of it. Miriam and Sarah told of gestures in Hebrew that survived and surmounted formal language gaps, sometimes *more* a matter of Hebrew interpretation than Dutch reality.

Bravo! Splendid provocations. "Make the familiar strange."

The numbers on a clock have no more intrinsic meaning than those on a road sign, no more intrinsic meaning than actors on a stage, than letters on a page. *We* breathe their life. *We* give them meaning, not of whole cloth, but of experience of lives lived, and not lived.

The classroom exists, ever fixed, ever changing, ever spawning eternal verities anew. Paula spoke of, "…reading 'disruption' through the lens of Brecht's methods…" Or perhaps expand it to: "…reading 'instruction' through the lens of *our* own curiosities, it is hoped we can begin to consider the possibilities for transforming the space in which we work so that we might mediate rather than dichotomize personal and public knowledge."

Yes, it all depends. And worse, it's indeterminate. We who educate children cast for third acts not yet dreamed. We start from experiential backgrounds beyond our ken. Realizing something of the indeterminate we teachers see time slipping by, becoming increasingly anxiety-ridden about precision-of-meaning and time-on-task. Ionesque gave us truth masked in

[3] Whitson, R., 2004. Reworking Place, Gender, and Power: Informal Work in Urban Argentina. Pennsylvania State University, doctoral dissertation.

[4] Salvio, P., 2017. *The Story-Takers.* University of Toronto Press.

the absurd. In education today our theatre *is* absurd. The clocks should read, “The time is now. And miles to go.”

1987

For a couple of years, I joined Lizanne Destefano and Deb Rugg as part of the college's National Transition Education Assistance Center, offering help with evaluation designs.

Programs that Work

At the Project Directors Annual Meeting (December 10-11, 1987, Washington, D.C.) Kathy McKean of Oklahoma's Project OVERS suggested that descriptions of successful projects should be bound and widely distributed. I agree that better ways of disseminating good work should be found, but I have strong reservations about present efforts of federal offices to disseminate information about successful models.

As noted by Frank Rusch and Bill Halloran, the U. S. Department of Education already has a Joint Dissemination Review Panel which reviews projects and introduces successful programs into the National Diffusion Network. The 13th edition (1987) of "Educational Programs that Work" containing write-ups of about 200 projects most recently surpassing the criteria of the reviewers. According to Bill, soon special attention is to be given to special education projects.

Unfortunately there is reason to question this evaluation and dissemination· effort, both in terms of identification of meritorious projects and as to the usefulness of the information to other educators. The criteria of success are heavily based on observable performance changes, such as improved student scores. Projects which choose outcomes measurement with this in mind are more likely to be certified.

I have yet to encounter a single educator who considers this source to be a valuable compendium. What it does is reward those who have invested great work in making a good program and then invested a great deal more in meeting the evaluation requirements. I admire programs that are more interested in investing that additional effort in further adapting the services to local need.

My second objection is to the notion that programs that work in one place could be counted on to work in many other places. Almost never has *generalizability* been included as a criterion in the design and operation. Knowing that successful generalizations are not to be expected, experienced practitioners seldom adopt whole programs, but steal bits and pieces from such sources to fit their own notions of what a good project is. So what the Diffusion Network needs is detailed description of previous program operations, staffing and the complex milieu. The name of the resulting publication should not be “Programs that Work” but perhaps “Programs that worked one place in some ways and perhaps are suggestive of a change here or there that other programs might make."

1988

Our evaluation work at CIRCE at the University of Illinois has been largely limited to program evaluation in education, sometimes with an emphasis on the curriculum, more often with an emphasis on program development. Created as the Center for Instructional Research and Curriculum Instruction in 1963 by Tom Hastings, with plans partly written by Lee Cronbach and Jack Easley, we of CIRCE regularly have been a group of three to four faculty members and six to eight graduate students, often collaborating with colleagues across the campus, but working more on off-campus programs than local.

At various times our staff has included -- besides Hastings, Easley, and myself: Ernie House, Gene Glass, Doug Sjogren, Jim Wardrop, Terry Denny, Arden Grotelueschen, Gordon Hoke, Claire Brown, Bob Linn, and Jim Raths, often just three on the faculty pay roll. But when Ernie joined in 1969, he brought along Tom Kerins, Steve Lapan, Norm Stenzel, and Joe Steele, swelling the ranks almost to double figures. Later, in 1977-78, we thought of the team working on the NIE-sponsored "Case Studies in Science Education" to be CIRCE "temps," a further swelling to include Mary Lee Smith, Lou Smith, Buddy Peshkin, Wayne Welch, Rob Walker, Rudy Serrano, Jacquie Burnett, Peg Steffensen, Beth Dawson and Chuck Secolsky. And with colleagues such as Klaus Witz, Del Harnisch, Liora Bresler and Mike Atkin; and with graduate students, such as Trey Coleman, Nick Smith, Barry McGaw, Bob Wolf, Deb Rugg, Craig Gjerde, Bob Louisell, Mel Hall, Aminata Maiga, Shameem Rakha, and Tom Grayson, we continued evaluating but with increasing emphasis on case studies.

Midlife Learning about Program Evaluation

Long ago there were pharmacies and apothecaries, ice cream parlors and general stores. Sometime about 1920 someone invented the drugstore. A newly registered pharmacist, my father worked in various drugstores in the early 1920s, then found himself with his own in Nebraska in 1925. Then, after 50 years of practice he sold his drugstore. A couple of years later, it was no longer a drugstore. And by and large all the drugstores had disappeared, transformed into pharmacies, supermarkets, and ice cream parlors. As a moment in all history, my father's professional life coincided with the epoch of "the drugstore."

Program evaluation had not been invented when I received my doctorate in 1958, but was, shortly thereafter. I am wondering if, like my father's, my professional life has spanned the beginning and end of my specialization. His was 50 years, mine, getting now to midlife, has been 25.

As I see it, my colleagues and I have become considerably wiser across the quarter-century, but seldom soon enough to provide just what our clients wanted. Though trying hard, we usually failed to teach them. All too often we patronized them. With roots in testing and instructional research, we regularly tried to introduce ideas and operations from outside education, often with humanistic consideration. But in a style well known to our colleagues across country and commonwealth, we failed to seriously engage the complexity and contextuality of educational practice.

Let me briefly identify three of our biggest learnings, these three regarding testing, politics, and epistemology.

Testing

Even within the first couple of years we came to the conclusion that for program evaluation purposes, testing doesn't take us far. When our clients deeply understood education they recognized that the tests on our shelves and in our catalogues did not deal effectively with their key objectives. Percentiles, yes, the predictive validities were fine, but the tests did not assess what students had learned, what teachers had taught.

Seldom were there time and money to develop more targeted tests. We joined briefly in the enthusiasm for criterion referenced testing, then came to believe that such tests were attuned to a stereotypical, simplistic interpretation of education and not around the complex, contextualistic expectations of our wiser and more experienced teachers and administrators.

But worse, we found that the understandings and decisions that one gains from evaluation studies were not sufficiently related to any student-performance criterion instruments to warrant a claim that those instruments had a validity for the declared purposes of evaluation. The sponsors wanted to know if their money had been well spent, if further support was warranted, if the teaching profession should be paying attention to the developments. Gain in student performance was too weak an *approximate* of the quality of a faculty's performance or of a contractor's performance.

Time and again we pondered Cronbach's[1] chapter in the 1971 edition of *Educational Measurement* reminding us that validity needs to be considered in a context of interpretation and use. Student performance data are important information to those responsible for the development of innovative programs, but we could not find justification for treating such data as program-effectiveness criterion data in most evaluative

[1] Cronbach, L. J., 1971. Test validation. In Robert Thorndike, editor, *Educational measurement,* Washington, DC: American Council on Education.

studies (even though a Request for Proposals might specifically define them so). Working over the years on such programs as Follow Through, we at CIRCE failed to convince Office of Education statisticians of the low relevance of regression on achievement tests scores. We did all too little with our doubts. Working out a theory of validity of evaluation studies, something well beyond the validity of the instruments, was a task we never got into.

Politics

One of our first major studies was directed by Ernie House,[2] a four-year study of the Illinois Gifted Education Program. Ernie and his colleagues became well acquainted with its operations, recognized its diversity and contextuality, and tried to compose the findings in highly conditional phrases. The state Associate Superintendent in charge objected to the "what-if" language, telling Ernie to write the findings simply and tell in straightforward language what needed to be done. He added that they would ignore our recommendations if they didn't like them.

And the National Science Foundation awarded us a very nice contract to help them blunt Congressional charges that NSF was out of touch with the American schools. Before our classroom case studies were completed, the protests cooled, and NSF (more or less unconsciously) redefined the project as a service to the American teacher. (Ernie[3] wrote a good book on the politics of evaluation.)

The National Science Foundation was satisfied with the case studies but what we thought worth learning was seldom what our clients wanted. They bought our studies mostly

[2] House, R. E., Steele, J. M., and Kerins, T., 1971. The gifted classroom. Urbana, IL: Center for Instructional Research and Curriculum Evaluation, University of Illinois.

[3] House, R. E., 1980, *Evaluating with Validity.* Beverly Hills, CA: Sage.

because it satisfied an obligation for them to contract for evaluation. We came to realize it. Some of us did not quit in disgust; we continued to enjoy the work and the elevated station, sometimes rationalizing that someone had to do such work, and that CIRCE probably might be less of a rip-off than others.

Epistemology

Most important, we feel we learned that anything worth being known by practitioners needs to be understood in its particular context, that the generalizations honored in the social sciences and by policy researchers are of little help in the classroom, or even the boardroom. We came to worry little about sampling, describing the action and cases in sufficient detail so that the readers could form a good idea as to their relevance. We came to do case studies, not for aesthetic, humanistic or political reasons, but for epistemological reasons -- they had meaning to the reader.

Increasingly we came to feel that weak education would not be made strong by us technologists, especially not by replacing subsystems or by implanting parts made outside. We did not come to be great admirers of American teachers, or their capacity for self-correction, but as with Churchill's view of democracy, “it beats the other ways.”[4] We came to see that the important knowledge for correction of classroom practice is experiential knowledge, and that the role of the evaluator can be to provide narrative accounts that provide vicarious experience.

It is clear that there will always be a need for evaluative judgment and rational problem solving. It is not clear that

[4] Churchill, W.S., 1947. Spoken before the House of Commons, 11 November.

Education needs specialists who call themselves program evaluators. Service to practitioners and provision of information will continue to be valued, as are registered pharmacists. Just what kind of shops we will keep remains to be determined.

1989

I became aware of the deeply-felt distress of my daughter-in-law, Kim Knutson, teaching in the Chicago Public Colleges, at finding that graduation rates were to be considered a major criterion of school quality. Her "English as a Second Language" students, intending only to attend, likely only being able to attend, a few courses, were, in a real way, being discriminated against.

Winning

In some ways Vince Lombardi, Coach of the Green Bay Packers, was an inspiration for athletics, but he did amateur athletics, and professional teaching, a considerable disservice with his emphasis on winning. Winning at any cost is wrong. A youth's education is corrupted if he or she believes that winning is the most, even one of the most, important things in life.

In National Collegiate Athletic Association narrative, at times seems as if graduation is the most important thing in college life. Overemphasis on graduation is as wrong as overemphasis on winning. Satisfying the requirements to stay eligible or to make minimum progress toward graduation are wrong notions of what education is all about. The NCAA should not contribute to that misrepresentation of Education.

Most philosophers of education would say that encountering and generating increasingly sophisticated ideas about the world, about communicating, about oneself, are more important than satisfying formal requirements. The requirements established by colleges are meant to assure that students will be gaining sophistication but students and professors alike know they are weak assurances. The NCAA would be wrong in ignoring college requirements but may be just as wrong in acting as if pursuit of graduation requirements is a necessary and sufficient condition for each student-athlete.

Each student needs mentors, academic advisors and a peer group intrigued by the issues of the various intellectual and social disciplines. Not infatuated but intrigued. Most universities leave it to students to find their own intellectual associates. Some of the smaller colleges work hard at assuring them. Good friends and instructors help a youngster orient to intellectual on-campus experience, and attend as well to preparation for careers. A few moment's thought reminds us that one can graduate from college without sound preparation for work, without intellectual commitment to lifelong learning, without awareness of civic responsibility, without appreciation for the greater scientific, artistic, and moral accomplishments of society. Graduation is only partial indication of educational success.

The responsibility for arranging good educational experiences for student athletes has been seriously examined by many coaches, athletic directors, and academic counselors. They know that graduation is not an ultimate, neither for the best students nor the poorest. They know that some students will not graduate, yet their lives can be enriched by the student experience. Not all student learning will be satisfying, enriching, and useful, but he or she can become better educated, more a good person, even in the few semesters they stay in school.

The academic quality of an athletics program is sometimes represented by the graduation rate of student-athletes. It is sometimes thought that any school that graduates a higher proportion of student athletes than other students has a healthy athletics program. We have sometimes derided programs that fail to graduate anybody. Such over-simplistic thinking may characterize Vince Lombardi. It is not a credit to the NCAA. What have we measured? Do we really know that the University of Chicago is doing more for its students than Malcolm X College? We have a lot to learn about the educational benefits of college for young men and women whose matriculation is marked most notably by dedication to athletic excellence, and to winning at any cost. We need to cheer other ways of representing the exercise of institutional responsibility.

1990

*My friend Saville Kushner asked me to write a frontispiece for his book, "A Musical Education: Innovation in the Conservatoire." He wrote about what we teach and why we teach it. I, trying to be amusing, insightful, and pertinent, this is what I wrote. I am pretty sure it failed to appear in the book Saville wrote. Nor is he in "*An Illinois Pair,"[1] *a case study I wrote for the Getty Trust about what we teach and why we teach it.*

Catching the Wry

The professional literature of Education is not a fountainhead of wry humor but one does find there the classic, *The Saber Tooth Curriculum.* To generations of teacher trainees, author Benjamin Harold[2] suggested that survival once depended upon escaping the bite of the saber tooth tiger. Fittingly, elders of family and tribe taught their young people various defenses. As teaching formalized around campfire and cave wall, they continued to teach avoidance of the saber tooth, even after such tigers became extinct.

Most curricula presume tomorrow's world will largely be today's. Most curricula presume we are capable of understanding today's world. Curricula obviously are structures of goals and objectives but less obviously are built upon platforms of presumption. As we should, we teach what we presume is needed. Often our motives are pure. Sometimes our presumptions are wrong.

[1] Stake, R. E., 1984. An Illinois pair: A case study of school art in Champaign and Decatur. In M. Day, E. Eisner, R. E. Stake, B. Wilson, & M. Wilson (Eds.), *Art history, art criticism, and art production* (pp. 4.1–4.58). Santa Monica, CA: Rand Corporation.

[2] Benjamin, H. W. R., 1939. *The Saber-Tooth Curriculum*. McGraw Hill.

To the naked eye, Education looks as fixed as the earth itself. In Education, much of Genesis seems still to be. Elsewhere, an information explosion, an electronics revolution; to the comers of the earth, "perestroika." Education alone, the steady state?

Appearances deceive. The earth rotates but we see a passing sun. The earth wears away but we plant again the same garden. Education itself consolidates, realigns, and wears away. Teaching and learning change, the purposes of education change, the people who control education change -- much of the change "obscure," lost in relative motion, below the surface, invisible to the public eye. Much of the change an atrophy-enervated public confidence, with few best hearts and brains to be found among teacher recruits, the schools themselves unable to stand up to presumptuous scrutiny of new-age monitoring, the media's substitution of sound-bite evidence of quality. Some of the change is reformative, some change repositions the constellations of knowledge in and about the universe.

In the beginning, teachers brought the best of themselves into teaching, not only what they knew and could do but what uniquely they cherished. Teachers now are selected and subsequently evaluated against a common template.

Students too. Part of what characterized reform education for a while a few years ago (some places) was an ideology of individual student uniqueness, with teachers responsible for fostering uniqueness as well as sharing common knowledge and skill. That leaning toward individualization has largely disappeared, almost without acknowledgement that it ever existed.

The teaching of the fine arts in schools is caught in the ordinary. No haven for evolutionary thinking nor fortress of individualism, a few arts education voices urging advance of the new or reverence for the old. Mostly the fine arts curriculum is one of our customs, Christmas songs, white paper snowflakes pasted on the windows.

What keeps it that way is seldom logic or need but politics. Parents, patrons, and prime ministers are enchanted with what has been or at least a notion of what has been. And the reasoning of artists and philosophers and eccentric teachers to do something else runs into a nostalgic illusion that what was *was* good. To be sure, parents should have a say in what schools will do to their children. To be sure, the society suffers by squelching its eccentric teachers.

Some geneticists lament a worldwide shrinking of the gene pool. To supply bakers of bread in Karachi and Caracas, nearby wheat farmers increasingly want the same hybrid seed that grows abundantly in Kansas. The native strains may contain genes resistant to new disease or adaptive to globally-warmed weather but agronomists lose access to those genes if the strains are not planted from time to time.

The creators of curricula seldom think about maximizing the grand store of knowledge in one of the best of granaries: the minds of children. For each and every child, they urge the currently most important knowledge. While here and there, a youth is devoted to the works of John Wesley, J. D. Salinger, Rene Magritte or another. Through its teachers, society has opportunity to extend or to quash diversity. It is not unreasonable to believe that dealing with the perversities of the future depends on nourishing diversity now.

Although prospects exceed those of the saber tooth, the fate of our species is in question. We do not know how much our survival depends on personal intelligence and collective knowledge. As always, we build on presumption. We may take a measure of comfort that human learning does not conform to the curriculum. Children learn much more than they are taught. Street corners, television and sports fields together deliver more education than schools. The sources of knowledge seem infinite, the channels of communication ever multiplying.

We are troubled as we should be by the quality of understandings gained through popular education. The current tactic of the schools is to make a few wry remarks about folk art and pop culture, not many, because it offends. Better to abstain, to ignore, to take responsibility only for respectable matter on which all children can be tested. Designing the curriculum is political. Not so much, as Marxists charged, to preserve capitalist structure and ethic, but to preserve the social standing of schools and the privileges of the profession. Individual professional educators remain remarkably dedicated to their epistemological disciplines and the presumed well-being of their students but the procedures for collectively arriving at a syllabus are shaped by faculty self-preservation.

In this milieu, we find the teaching of the arts. Not unlike those for the teaching of science and history, arts curricula are laid out as to what is best for all. Interpretation is greatly loosened because the general public has little desire that the arts be part of general education. Here and there, schools are to provide a small contact with the arts. A sample should be accessible, perhaps briefly enriching the experience of children for whom one or more art forms will be life-involving. As far as people outside schools are concerned, what is taught in the arts is not very important. Some program should be maintained and certain exhibits and performances perhaps continued.

And thus the opportunity is great! Everywhere else the curriculum is bogged down in historical precedent, preparation for college, notions of minimum competency, preservation of the status quo. Outside the arts, no one much cares what is taught as art. Those who teach the arts have less resource than those down the hallway, but more freedom. What they choose to vivify is what they hold dear, madrigals or graphic design or

Morris dancing. To society's diversities are added the teachings of many an artist and arts teacher.

The opportunity for diversity and special experience exists, often ignored. To many arts educators, the arts are seen as needing to be elevated to the status of the other subject matters, therefore standardized and disciplined, postponing excursions by individual students until pre-arranged introductory exercises are completed, and the elements mastered.

It need not be that way. School art is the embodiment of promise. Most people of the world have an idealism about Education, an idealism hobbled by the idea that children be trained before they be educated. School art now holds little priority in the grand sweep of things. Thus it is privileged with under-expectation, with little demand for mass-produced products. Some artists are aware of this richness within poverty, more educators might heed.

The preservation of fine and popular arts alike probably does not depend on storing copies of the masterpieces in deep caves. Perhaps it depends on maintaining a society sensitive to change, admiring of diversity, tolerant of what seems awry, willing to poke a little fun. After all, creativity is too important to leave to the Creator.

1992

This one was prepared for the annual meeting of the American Evaluation Association, Seattle, November 1992.

Presumptuous Indicators

Up and down these Sheraton halls, I catch the presumption that because our findings follow data, they are credible. I think our work is filled with presumptuous leaps, sometimes disciplined, often not. We produce student achievement scores and interpret them as indicators of teaching effectiveness. We gather user satisfaction comments and allude to program effectiveness. Sometimes we could out-misrepresent Madison Avenue. Raising standards of evaluation performance includes reining in our presumptuousness.

It isn't that we gather too little data. It isn't that we should label those findings that are data driven and those that are not. Many of the interpretations we make appear in paraphrase, amplified illustration, and advice. Some of these extrapolations are valid; some go beyond what has been validated. I think we should work to assure ourselves that our presumptuousness is tempered by scrutiny and skepticism.

Perhaps we actually collect too many data. A lot of them serve merely to "legitimize" our work, adding little to the understanding of the meanings and quality of our evaluands. We should probably spend more time dwelling on the meanings of our key data, checking the limits of our generalizations; more time with final report review panels; maybe a larger portion of AEA meetings seriously critiquing our findings.

I do not mean to suggest that the few people who pay close attention to our final reports have been pressing for less presumptuousness. Some care little. Regardless of how carefully we might validate our findings, many people use them to back up the claims that suit them. In 1979, a spokesperson for the widely respected American Association for the Advancement of Science ignored our explicit finding that there was no national demand or willingness to use innovative science curricula. He claimed the opposite, that our study showed a need for a new round of curriculum development. Beyond a few cautionary statements, it isn't the job of evaluators to improve rationality and cautiousness in our clients and audiences. It is important to be realistic about the world we work in, but we should not public use be our standard.

We should raise our standards of interpretation. I think we have become a part of "the problem," rather than part of "the solution." For me, the center of the problem is our *scientistic* evaluation system. We pressure our readers to accept simplistic representations of the quality of complex offerings. We represent student understanding with a score on a standardized achievement test. We represent vocational training success by graduates employed. We seem to have lost the appetite for comprehensive validation. Many of us haven't looked at Campbell-Stanley threats[1] since final orals. It is not apparent to most people how presumptuous we are.

Except maybe to the young. Have you noticed that we are not getting a fair share of able young people to enter the evaluation field? Quite a few students take a course or two but few prepare for a career. There is plenty of evaluation work to be done. I find that many students see us as corporate and bureaucratic voices. If we were more careful about our indicators, we might attract more apprentices.

The bad guys here are us who describe educational phenomena with indicator systems. Of course, we have to have

[1] American Psychological Association, 1999. *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

indicators. (Some more accurately call them “approximates”.) All science, all communication, is made up of indicators. Words are indicators. Measurements are indicators. We couldn't think without indicators. But we could care more about validity. In Education, one gross example of carelessness was the 1983 posting of the “Wall Chart” by U. S. Secretary of Education Terrell Bell,[2] implying that a few thousand SAT scores from a nonrandom sample of students represented the quality of Education in each state. Admitting his chart had little validity, Bell merely said, "How could we get something better?" I say, "How better to trash our profession?"

On the question of validity of our measurements of education, the *Joint Standards for Educational and Psychological Tests*[3] have set high standards. They call for strong evidence, but only for a small part of our work. We often do not speak carefully about our measuring. We ignore the need for cautious interpretation. We allow too many misleading representations in our papers, our presentations, our conversations, our consultations, our teaching. We should continue to speculate, to approximate, but we should draw a clearer line between speculation and finding.

[2] Bell, T., 1983. *A Nation at Risk*. National Commission on Excellence in Education. Campbell, D. T. and Stanley, J. C., 1963. *Experimental and quasi-experimental designs for research*. Chicago IL: Rand McNally.

[3] American Educational Research Association, 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

1992

For seven years CIRCE had a year-to-year contract with the Chicago Teachers Academy for Mathematics and Science to evaluate efforts to improve teaching in schools that volunteered for the services. We tried to spend as much time as we could in participating-teachers' classrooms.

Here is part of an evaluation case report on Harper School, ten days on site, allowing only cursory study of its teaching and administration. In other rooms, I observed what I perceived to be sometimes strong and sometimes ineffective instruction. The principal's remarks confirmed a considerable range. It was my judgment that teacher competence, (the TAMS shtick) however imperfect, was not a major obstacle in TAMS schools to raising student achievement.

Shadow Study of a Sixth Grader

At 8:30 a.m. on Thursday morning, Adam shows up at the cafeteria door. Breakfast is being served but Adam doesn't go in. The woman giving out meal chits has her hands on him, seems to be sparring with him, verbally. And then he disappears. Adam is one of five siblings, all arrive at school in the morning with less than usual parent attention. Short, with a beautifully sculpted head and Gerri-curl, solid body, baggy black sweats and sneakers, and full of energy, Adam is a person of notice.

At 8:55 he climbs the stairs to the third floor with other upper graders, turning to block the girls behind him and thus a string of others. Adam manages to keep the girls off-balance until Ms. Crain, one of the teachers, spots him and gets traffic moving again.

The "augmented staffing room," for Adam and 15 other "at risk" children, is at the top of the stairs. It's Mr. Garson's fifth-sixth-grade room. Garson notices Adam, has a few quiet words with him before a paternal shove toward the room. Adam disappears into the closet and sheds his oversize coat. Garson tells him to get the dust mop and clean the floor near his desk, the closest to the door. A dozen or so children are on time, milling about. Adam gets the mop and runs it into the feet of those nearest. The hurly-burly of the hallway is dying away and the children within this room are becoming quiet too. Intercom: "Ms. Hampton, please come to the office."

It's a typical elementary school room with full windows on one side, blackboards across the front, homemade and purchased posters almost everywhere. Near the door, two twigs of cotton are labeled "Cotton." Movable desks are clustered to each side of a commons, eight on each side, most facing the blackboard. It's a big room for so few youngsters.

"Excuse me! Excuse me!" It's Garson's way of quieting the class. A few students whisper and Garson works at them to get things straightened out. He has briefed me how important it is to establish discipline. At 9:10, two tardy children arrive. He asks for their excuse. The girl was getting a younger sister to school. The boy ignores the question and is sent to the hall "to straighten your face." It becomes apparent to me the class is waiting for opening ceremonies via intercom. Adam sits quietly at his desk.

Garson announces, "DuSable High School Band and Gospel Choir will be here today" and mentions Tina Turner. Although he has taught in the U.S. many years, I have difficulty understanding his Caribbean dialect. The students show no sign of not understanding him. Perhaps Garson is asking who will sing. Adam shows no interest.

With quiet again, Garson grumbles about the lost time. "Now, how many want lunch?" Only Annalee raises her hand.

One by one, he checks each child, bantering, loosening up the kids, getting away from the artificiality of the announcements. "How about a hot dog, Adam?" "Naw, burger 'n fries." "Well, you'll have to go somewhere else for those. Adam, would you sit at your own desk!" He does not.

"Darci, would you read 'Strange Things in Space'?" Her second sentence refers to junk food. Garson breaks in. "What is junk food?" No one answers. "Candy and cookies and anything not the regular food we're supposed to eat." Darci reads on. When she encounters a difficult word, Garson supplies it. And he pushes the rest to interpret what they are reading. "What do they mean, space junk?" The intercom: "Teachers, reminding you to send your absence slips to the office." "Millions of stars and what else? There is even a special star you can't see at night. And shooting stars." Darci finishes the paragraph. "You can learn quite a bit by looking at stars. You can make a map of them."

"Okay, Adam, you read." Adam reads, "The earth moves around the sun." And onward, words frequently supplied by Garson but reading at good pace and with a sense of sentence. When he reaches the use of binoculars and telescopes, Garson takes over and tries to draw the students into prediction of what they will see with binoculars and telescopes. "Yes, they make the stars look larger." He wants a prediction also of what space junk they will see. "At a planetarium -- you've all been there -- you can see how the stars move in the sky. And their size and color. What did Adam read about the name we call people who spend their lives studying the stars? Astronomers. Astronomers use very large telescopes." Everyone is quiet. "Say it, students. I want just the smart students to say it." "Astronomers." (Pause.) "Thank you, Adam."

Kim reads but is soon interrupted. "Why are the stars seen in different places every night? Yes, they move around. What is the key word here? They move, don't they?" And after something about an observatory at the top of a mountain: "Why is the air clear atop the mountain? Who says because it's so high? That's right. Closer to the ground you have all that stuff in the atmosphere. Kerstin put your hands down. Why do we have falling stars? Come on. You have to know that, you who'll be seventh graders next year. Because they are old? No. Do you think they just say, 'I'm tired of you, Mama, gonna fall now.' Think." And Kerstin reads some more. "And why is the sun yellow? And not blue or green. Make some suggestions. Okay, look it up in the book. Page 170 tells you why stars fall." And then he goes into something about the song "Catch a Falling Star" and how Adam might write a Valentine to his girl and want to send her a falling star. Kim reads aloud again. "And on Page 172 you'll find space junk: dust and rocks. Now you need to know the difference between meteor and meteorite. Work on those definitions."

Adam is back at his own desk. Annalee reads again. The class is following along. All turn the page together. Mr. Garson helps Adam keep oriented to his book, maybe partly because Garson was forewarned I am trying to keep a hidden eye on Adam. It is 9:55. "Put your books away. Straighten your desks around." Adam continues to read. "Get in line." Adam stands talking to Garson turning away but gradually (actually) leaning against him.

All off to the washrooms, Mr. Garson ahead, Adam trailing last. They pass a line of eighth-grade boys. Adam succeeds in drawing their eyes, turns to them, thrusts his pelvis, covers genitals with both hands, and says something of which I only hear ". . . pussy." A huge smile on his race, a few smirks on theirs.

The Black Heritage music program gets a fine response from the several hundred students. Adam, however, sits 90 minutes without expression, mostly with arms folded. On the way back, during the washroom stop, another teacher grabs Adam and reprimands him. Mr. Garson takes Adam into the corner and talks to him several minutes, then puts him at the room's Apple terminal. He acquiesces, works attentively while the other children study English until noon. The chapter theme is Winter Weather. Adam misses the orientation to "topic sentences" and the class discussion about stuffing broken windows at home with paper and plastic and taping bullet holes in project windows to keep out the winter wind.

Detail of the afternoon more or less mirrored that of the morning. Adam avoided submission, susceptible to special discipline, always a potential disruption to lessons. On the way to my car I thought: Adam is not a typical youngster and Mr. Garson not a typical teacher and this perhaps was not a typical day (and mine may not be typical eyes). Yet it is apparent that this teacher had a major task in working with the personality and the needs of the boy -- and other teachers have similar unending responsibility for the socialization of children and the restoration of a few hobbled lives. Mr. Garson had a small and responsive group to work with, including Adam. Adam was responsive. He seemed not to lack aspiration or self-esteem. Somehow he appeared to me to be a person who would impact people's lives. What would be his mission? Today it seemed far from the school's mission.

Like that of many other teachers, Mr. Garson's approach was confrontational. Some of the time it required confrontation with students and regularly it confronted the demands of the community, the children, the school, and the school reform effort. Cultivating the academic talents of kids was important in Mr. Garson's room, but responding to their personal and social needs apparently often took precedence.

1994

Even then, Tom Hastings and I were "testing persons." The third member of CIRCE's original crew was Jack Easley, a "curriculum person." He too looked for quality in teaching and learning, and he wanted to see it with his own eyes.

Jack Easley, Insight Finder

When Jack returned from surgery in Cleveland, he went back to work on the *Science Network News*, his newsletter for teachers and children, answering questions that kids asked. He died a few days later. Mindy Miron, our CIRCE office helper, put the finishing touches on the issue Jack was working on. The last kid questions he tackled were:

What did God do before He created the Universe?
Do Christmas tree lights that blink take more electricity?

For all his professional life, Jack worked on getting teachers to talk about the science ideas that kids have. That is why he and Bernadine started the DIME group, Dialogues in Mathematics Education, and part of the reason why, in the 70s, he and Klaus Witz and Jacquie Hill started their Committee on Culture and Cognition. They were of the belief that educational reform, Max Beberman and Don Bitzer and Dick Suchman notwithstanding, that education reform was more a matter of upgrading conversations in the classroom than in upgrading the delivery of mathematical and logical paraphernalia.

Sixty years ago, in the long, war-year winter of 1944, Jack would walk alone across the frozen ground near his camp on Baffin Island, watching the stars, pondering questions of the universe.

In his last years, I often thought of Jack as "the answer man." On my radio sixty years earlier there was the Answer Man who knew answers to the most difficult questions. Jack didn't know the answers to most of the questions kids would ask but he knew that lots of people have lots of different answers, and part of the essential study of mathematics and science is getting answers talked about -- particularly the answer of each child -- talked about, right or wrong, talked about. He sometimes called it, "minds-on" science.

I paused to reflect when JoAnne Fley read me her draft of Jack's obituary which mentioned his work as a program evaluator. Jack talked very little about the theory and practice of evaluation and probably never attended a meeting of the American Evaluation Association. But he was an evaluation specialist before there were evaluation specialists, and 25 years before an American Evaluation Association was formed.

Max Beberman, the Illinois mathematician and one of the founders of the "new math," came across this Harvard-educated assistant professor in Honolulu, and brought him and Elizabeth to Urbana in 1962 for Jack to be the internal evaluator for the UICSM math project. Jack watched Master-teacher Max, then listened to the problems that the UICSM textbook writers faced:

Should they insert an example here before the explanation or after?

How should they use the concept of sets to explain correlation?

What would the students think?

Jack was proud of himself by so lubricating his feedback mechanisms, he loved the word "mechanisms," -- so proud of his feedback mechanisms that could get student responses back from pilot classrooms to the writers in two weeks. It had been well more than a month at first. But even in two weeks, disappointing, to find that the questions had changed. His feedback to the writers was for last week's questions. How was one to deal with the immediate!

At a meeting of the Social Sciences Education Consortium over at Purdue one time, Lee Cronbach talked about Jack's work and caught the attention of philosopher Michael Scriven. A couple of years later, Michael more or less invented the specialization of educational evaluation, laying down many key definitions, but one most remembered, that of formative and summative evaluation. Formative was the very evaluation that Jack was doing and the very thing that Cronbach emphasized in his views of evaluation. With Tom Hastings and Lee, Jack wrote the proposal that created CIRCE in 1963, and helped Tom staff our evaluation center with me, Ernie House, Gene Glass, Doug Sjogren, Terry Denny, Gordon Hoke and Arden Grotelueschen, and ultimately Claire Brown, Harry Broudy, Jim Raths and Bob Linn. And why would we not count as staff the constant stream of stay-awhile visitors: Clem Adelman, Sid Dunn, Larry Ingvarson, Gary Joselyn, Don Hogben, Helen Simons, Mike Bikalis, Sue McBurney, Ulf Lundgren, David Hamilton, Barry McDonald, and Saville Kushner, to name a few.

You probably couldn't get Tom -- or Chip or Susan Bruce -- to say so but Jack was terrible to write with. He couldn't let an idea sit still long enough to really write about it. He would write a page one day, a page we all admired and were ready to elaborate into the grand theme, but soon he would redevelop the idea, an even better idea -- and, with every draft, it happened anew. Not that there weren't wonderful things then to talk about, but getting that proposal in the mail or writing that final report was like catching the North Atlantic stars.

Formative evaluation for Beberman's writers turned for Jack into a paper with Russ Zwoyer, *Teaching by Listening*[1], then into the qualitative study of teachers and children, especially with Klaus and Jacquie, then for 3 years with me into our NSF project *Case Studies in Science Education*, then for several years with Bernadine video-taping in Susan Shadid's classroom in Kankakee: ordinary children thinking extraordinary thoughts. Elizabeth worked all this time with Jack on some of his deepest penetrations into "minds-on" science.

Jack was a listener. He heard much of what a person was trying to say. To a Jean Piaget. To a fourth grader. To a Ros Driver. To Bob Louisell and many another graduate student. Was ever an idea wrong? Was ever an explanation boring? ... for each had a facet, a twist, that made it original, a cause for pondering, an unfolding of something maybe really important. One of the best parts of a DIME meeting was Jack's expression of appreciation for what few of the rest of us had managed to hear. Even the night sky talked to Jack.

A kind of one.

[1] Easley, J.E. & Zwoyer, R.E., 1975. Teaching by Listening – Toward a New Day in Math Classes. *Contemporary Education* (47)1: 19.

1997

A two-month evaluation of the Reader Focused Writing *training of the U.S. Veterans Benefits Administration was taken up by a CIRCE team of Rita Davis, Stephen Glynn, Kathryn Sloane and myself.*[1] *William Platt was the VA coordinator. Congress was regularly getting complaints from armed forces veterans whose application for support for medical, housing and education, had been denied or were less than fully supported. Too many applicants said the letters they got back were incomprehensible. VA set up an on-line training program to coach all department letter writers, hundreds of them, in better letter writing.*

Evaluation of "Reader Focused Writing" for the U. S. Veterans Benefits Administration

During the period, August 1-October 31, 1997, a team of evaluation specialists from CIRCE at the University of Illinois evaluated the Fall, 1996 VBA staff training in "Reader Focused Writing." Five Regional Offices were visited, a national survey was administered to directors and trainees, and letters to veterans were examined pre- and post- training. A meta-evaluation study culminated in a three-hour hearing on October 29 (three months after the start) under the direction of Vice Chancellor Stephen Kemmis of the University of Ballarat. A draft of the CIRCE report was submitted to the Veterans Benefits

[1] Stake, R. E., & Davis, R., 1999. Summary of evaluation of "Reader Focused Writing" for the Veterans Benefits Administration. *American Journal of Evaluation, 20*, 2, 323-344.

Administration on November 6.[2] A summary of conclusions follows:

Evaluation of Need, Goals, and Plan

Under the title of Reader Focused Writing, a coordinated effort to upgrade written communication within the Veterans Benefits Administration resulted in the Fall, 1996 training of 775 staff members at Regional and Central Office sites. For each trainee, the training lasted 21 hours and, at the regional sites, was carried on via an interactive satellite network. Presentations and activities centered on increasing attention to the veteran’s needs and making communication comprehensible and useful to the veteran.

Appeals for better letter-writing had been long heard from veterans, legislators, the President, GAO evaluators, and from VA staff persons themselves. Some criticisms overstated the problem, but a problem existed. Despite remedial efforts over the years, the need for more considerate and useful letters to veterans remained. Studying a sample of letters sent out by Regional Offices, our panelists concluded that much more should be done to consistently reach high standards of communication. The high majority of letters were at least satisfactory but a troublesome minority remained unacceptable.

The goals of the RFW program were to upgrade communication awareness and effort of the entire VBA workforce. In a corporate sense, all workers are engaged in communication with veterans and it behooves all to understand the problems and possible alleviations. The instructional goals of Reader Focused Writing were clearly appropriate for the VBA.

But it was clear that getting everyone trained, or getting even the subgroups primarily responsible for letter-writing

[2] Stake, R. E., Davis, R. & Guynn, S., 1997. Evaluation of Reader Focused Writing.

better trained, would not by itself eliminate faulty communication. There were impediments to communication beyond insensitivity and low capability. One impediment was -- with an increasing flow of letters from a staff being downsized -- an institutional need for high productivity. Another was the lack of consistent support from Agency leaders for re-engineering the writing. The expressed goals for RFW were clear and legitimate but were not given sufficient priority Agency-wide to dramatically change actual practice.

The need for staff training was real, but equally real, perhaps even more pressing, was the need for getting PCGL and other computerized letters fully compatible with Reader Focused Writing. RFW principles of communication are sound principles for all communication. Many of the PCGL passages continued to be obscure, impersonal, and traced with defensiveness. This is not to say that all claim resolutions can be stated in a way every veteran can understand, but complexity is not the major problem. The major problem is one of priority. Good letters, whether originating as pattern letters or not, take longer. To best resolve the conflict between "timeliness" and "veteran oriented writing," it is important to have a staff keenly aware of good and bad letters. The plan for *general education* of the staff in RFW was responsive to the need and well thought out.

Impact of the Training

We found indications that a high majority of VBA staffers were persuaded even before RFW training that their letters to veterans were low on sympathy and high on technicality. The training reinforced the desire to do better, but it gave them many unexpected ideas about how to be more considerate of the veteran and to deliver a more useful package of information. Quite a few trainees reported observing some changes in

Regional Office operations attributable to the training. The most visible impact was to move them into appreciation of the RFW format, with headings, bullets and white space. But the more powerful impact was to persuade them to worry less about being precise, complete, and protective if it might make the letter more comprehensible. According to testimony, they came away from the training with greater understanding of how a letter is read. And according to our analysis of the letters they authored in our contrived performance situation, they were able to put into practice the principles of RFW. And our panelists reviewing RO letters which were offered us from the file, even those from PCGL software, found a few more high-quality letters after the training.

But we had little reason to conclude there has been change in quality in the volume of letters actually mailed to veterans each day. Regional Office personnel did not expect a change because PCGL structures are changing too slowly and because the workload for Adjudicators and others who use the structures is too demanding to allow customizing or even to carefully proofread the work that they do. They know how to write better. The system does not require them or allow them to create and guarantee the best of which they are capable.

Quality of the Training Method

Interactive satellite instruction was used during the Fall 1996 Orientation and Tools Course for two hours of training each day followed by an hour of small group problem solving. Some found this running on too long. On the big screen, presentations were informal but highly accomplished, engaging the eight or ten staff members gathered at each site, attentive to questions raised and closely tracking a well-developed instructional plan. It was well done. Essentially all of the data we gathered were based on instruction under a single impresario, Melodee Mercer -- which means that we do not know how effective the training would have been under a less talented leader. Continuous emphasis was given to orienting the letters to the veteran, anticipating his or her needs and laying out clear steps to be followed or options to be selected.

Distance education (that's what this had been called in Education circles) in industrial and scholastic settings has often been ineffective because, even more than in the classroom and auditorium, the audience remains disengaged from the learning, sometimes hostile to it. The interactive electronic equipment used with RFW, allowing voice contact between all trainees and the instructor (although visual imaging was only of the instructor), did not function well much of the time, but the trainees at most stations were regularly and comfortably engaged. In some cases, the engagement was facilitated by the on-site instructor, but most of the attentiveness was drawn by the quality of ideas raised by the lead instructor, the instructional materials, and particularly the illustrative letters to and from veterans. From all our evidence, it was apparent that the participants found what was being taught worth knowing.

The expectation of many was that they were there to learn how to write better letters, and the lessons did not dissuade them. Over and over, the training beamed in on writing original letters to veterans. But this RFW Tools training really was not justified as skill development. Half the trainees had "letter writing to veterans" as less than 20% of their total responsibility. And they and the other half were locked into computerized letters. Almost all VBA letters to veterans were developed on PCGL software or the equivalent. Staffers were not allowed the extra time to write more personalized letters. Some trainees spoke of a mismatch between the training and their work. But all personnel did need to understand the RFW system. The instruction could better have acknowledged wide differences in work responsibility across the Agency and given more attention to the central role of PCGL but teaching about Agency responsibility for effective writing by concentrating on the writing of original letters was not necessarily a bad strategy.

Impact of VBA Infrastructure on RFW

RFW training did not persuade those who controlled the workflow in the Regional Offices that a better compromise between productivity and comprehensibility should -- or perhaps could -- be achieved. Productivity was a higher standard. Many stations were often faced with a backlog of files, turn-arounds extending into the months, unable to get the job done even with the absolutely quickest of answers required by law. Staffers did not expect their bosses to give them opportunity to write a more compassionate, a more useful, letter. VBA is a top-down organization, the workers do not vote on performance standards. Training the workers to recognize and want better letter writing is no more than a wisp of influence on how the VBA will be run.

It is possible that under a different calculation of productivity, one taking into account the extra correspondence required by veteran requests for explanation and mis-response to VBA requests, a better balance between productivity and comprehensibility might be reached.

The ostensible fact is that the Regional Offices supported RFW. At perhaps half the stations there was surprisingly little reluctance to have a large fraction of the staff participate in the training. Concerns about the quality of VBA letter writing appeared deeply felt, but administrators implemented neither the rewards for good letters nor the censure for bad that might have changed things. The principles of RFW communication were endorsed. With some apology, Directors spoke of their inability to find a way to answer the mail in timely fashion and, at the same time, adhere to expectations codified by RFW.

We found ample disbelief in the Regional Offices that the Central Office supported RFW. They pointed to C.O. communications reflecting little awareness of RFW. They saw little C.O. encouragement for mitigating timeliness standards in order to assure that each outgoing letter was carefully re-read. They knew that some people at the Central Office as well as Regional Offices had worked hard on the Task Force and on creating the training package but that they had less than full support across the Agency. The infrastructure for Reader Focused Writing was infirm at both federal and regional levels. Future extension of good RFW training could not be expected to generate, by itself, the climate essential for improved letter writing.

Grounds for Continuation of RFW

The general reaction of Regional Office people to the discontinuation of RFW telecasts in November, 1996 was disappointment. Those who had participated found the sessions informative, persuasive, and a good opportunity to extend their understanding of their Agency. Even if obstacles to incorporating RFW into their work were formidable, they felt that the training enhanced their overview of responsibility. They felt, with good reason, that others should have opportunity of taking the Tools course and that more advanced work should be offered. Nor did there seem reason to discourage the supplementing of RFW with ordinary courses in grammar, business letter writing, and composition.

In fact, one of the more important side views of this evaluation study was that the Regional Offices appeared generally in need of a professional development environment or ethic. In-service education should not only be keeping up with technological change but enhancement of the worker as a person with professional responsibilities. The "Reinvention" process developed at the New York Regional Office, for example, indicates that staff members will be "implementing the design somewhat differently in different teams: some team's case managers do more of their own development than others." Other stations are moving in the same way. Such decentralization of responsibility requires an approach to training based partly on what is good for the individual worker, a professional development ethic. The interactivity and small group sessions of RFW training were consistent with this ethic.

The aims of the RFW program are likely to be best served by extended effort to upgrade PCGL-type holdings. Formatting and logic in this software are highly sophisticated but, over all, falling short of the personal view and utility standards advanced by RFW. Supportive of this effort should be the forthcoming RFW guide or reference book to help letter writers deal with unfamiliar problems. Future RFW presentations should be able to show letter writers using PCGL, working through unfamiliar problems, without diminishing emphasis on its principles of communication. Bringing in telephone communication will also be a needed extension of RFW. Future use of RFW should tie in with present job analyses of VBA personnel which, of course, will emphasize PCGL structures and voice interaction with veterans.

Reader Focused Writing is a major asset for the Veterans Benefits Administration. It is built upon sound principles, has the endorsement of the directorate and widespread acceptance among rank and file across the country. The training activities can be improved but, by and large, they have had a strong positive impact and appear well worth the investment. RFW addressed the recognized need for better communications.

1999

Widespread interest in measuring the effects of social and educational programs created a pressure on evaluation designers to overpromise what they can measure and to underemphasize reporting of meritorious conditions as defensible purchases from program investments. I commented on contemporary views.

The Misanthropy of Effect

1. The work of evaluation, including program evaluation, is seen by many as the determination of the effect of an evaluand. One of the primary reasons that clients contract for evaluation services is to determine outcomes, that is, the effects of the evaluand.

State of the art

2. For many programs or program changes, given contemporary technology, the effects cannot be measured directly. Indirect measurements of effects, also called *approximates* and *indicators,* are often based on mere appearances, aka *face validity*,
rather than on validation studies.

3. Anticipating emphasis on effects plus lack of caution in the interpretation of indirect measures, evaluators have knowingly promised more in their proposed contracts than they can deliver.

Cause and effect

4. For many programs or program changes, even when we can measure a certain effect, it is difficult to attribute it, even in part, to the evaluand as the cause.

5. The conceptualization of cause and effect is a central part of social science but it is not an essential feature of disciplined study of social processes. Gathering data on happenings or processes may be more useful.

Artificial significance

6. Many evaluation theorists and practitioners are satisfied that a causal relationship is established if it meets a conventional statistical standard, usually some measure of group difference or correlation.

7. An effect that changed an entity by a millionth part, were it to be measured accurately over a million cases, would be found statistically significant even if it were an effect not humanly discerned or useful.

8. Social policy should be more attentive to human discernment and valuing than to statistical significance. An over-emphasis on effect in program evaluation diminishes the recognition of conditions widely considered meritorious.

2000

I intended to -- but may not have -- submitted this to Educational Policy Administration Archives. *I don't believe it was published.*

We did not know much about what assessment was accomplishing but we knew it had not brought about the reform of American Education. The costs and benefits of large-scale mandated achievement testing are too complex to be known. (Economists are always willing to guess.) I argued here that educational policy needed to be based more on locally deliberated interpretations of assessment, experience, and ideology. Evaluation of assessment consequences, however inconclusive, needed to play an important role in the deliberations.

Assessment in U. S. Education

During the last half of the Twentieth Century in America, the traditional quality-control of schooling, i.e., informal management (by teachers as well as administrators) Board oversight, parent expression, state guideline and regional accreditation, have continued to be prominent in school operations. But because the perceived quality of public education has fallen off, other means have been added to evaluate and to improve teaching and learning. For thirty years, formal assessment has been a significant means of quality control and an instrument of educational reform.

In last Century's third quarter, 1950-1975, the impetus for changing American schooling was the appearance of Sputnik. It was reasoned that American schools were unsuccessful if the Soviets could be first to launch spacecraft.

College professors and the National Science Foundation stepped forward to redefine mathematics education and the rest of the curriculum, creating a "new math," inquiry teaching, and many courses strange to the taste of most teachers and parents. According to Gallup polls year after year, citizens expressed confidence in the local school but increasingly worried about the national system. In the 1960s, curriculum redevelopment was the main instrument of reform but, in the 1970s, state-level politicians, reading the public as unhappy both with tradition and federalized reform, created a reform of their own. Their reform spotlighted assessment of student performance.

The term "assessment" then became taken to mean the testing of student achievement with standardized instruments. Student performance goals were made more explicit so that testing could be more precisely focused, and efforts were made to align curricula with the testing. Schooling includes many performances, provisions, and relationships which could be assessed but attention came down predominantly on the students: "If they haven't learned, they haven't been taught," was the cry.

Now for at least two decades, in almost every school, at every grade level and in each of the subject matters, student achievement has been assessed. And every year, it has been found largely unchanged from previous testing. Over the same periods, teaching, on the whole, appears to have been little changed, certainly not restructured. Explication of goals appears not to have set more achievable targets. The last decade has seen efforts to set standards particularly for levels of student performance needed to restore American Education to a leading world-position. From time to time, gains occurred, but small and not sustained. Losses also occurred. Instead of reading this lack of sustained progress as pointing to need for a different grand strategy, the clearest summons has been for additional assessment.

Purposes and expectations of assessment

Goal statements are simplifications. The purposes of Education, aggregated across the profession, across researchers, the public and the primary beneficiaries, are far more complex than those represented in goal statements and formal assessments. Facts, theories, and reasoning are needed not just in isolation but interactively, innovatively, in a range of contexts. We hold a vast inventory of expectations, beyond catalogue, partly ineffable, often only apparent in disappointments as students appear poorly taught. That immense inventory is approximated by the *informal* assessments by teachers much better than by explicated lists of goals.

The grand manifold of purposes of Education held by any one person at any one time also is complex, and situational and internally contradictory. People, even those specially trained, are not very good at speaking of "what all they expect" of an educated person. Again, the complexity shows most forcefully when the person does not perform well. Any one shortfall tells little about the array of purposes. Any one assessment, however precise and valid, does not sample well the manifold of purposes. Broad and attentive use of assessments, formal and informal, evokes realization that what we expect of students and the uses to be made of a graduate's education extend far beyond formal goals, standards and lesson plans. Formal representations of aim and accomplishment provide flimsy accounts of the real thing.

This is not to suggest it is useless to record educational purposes and student performance. It is useful to categorize them, to illustrate and prioritize them, sometimes by abilities and subject matters -- but always a risk. The subsets and domains are artificial. Needed in the anticipation and provision of Education, they often serve poorly to represent the education a student is attaining. Assessment based strongly on goals and domains is likely to tell more about the territory of teaching than the territory of learning.

Procedurally, Education is organized at the level of courses and classrooms, then lessons and assessments. Actually, education occurs in complex and differentiated ways in each child's mind. Assessments tuned to management levels cannot be expected to mirror the complexity of learning and diversity of learners. However carefully named and designed, mean scores do not necessarily indicate basic accomplishments for a group of learners. Each testing needs empirical validation.

Validation of assessment

Standardized test development is one of the most technically sophisticated specialties within Education. Definitions and analytic procedures, at least at the major testing companies are scrutinized, verified, codified and reworked. The traditional ethics of psychometrics call for extensive construct validation of the measurements to be used in schooling. And it is not enough that the instruments and operations be examined for accuracy, relevance and freedom from bias, but that independent measurements be used to confirm that scores indicate what we think they indicate. Sound test development is a slow and expensive procedure.

In the development of assessment instruments by the 50 states, adequate validation has seldom taken place. Instruments have been analyzed statistically to see that they are internally consistent but not that they mean what users think they mean. Presumption that assessments indicate quality of teaching, appropriateness of curricula, and progress of the reform movement -- commonplace presumptions in political and media dialogue -- is unwarranted. Proper validation would tell us the strength or weakness of our conclusions about student accomplishment. Such studies have not been commissioned. The most needed validation of statewide assessment programs has not taken place.

The question of whether or not the assessment legislation, as opposed to the assessment scores, is having a good effect on student education is a separate question. Assessment changes instruction. Reformists expect assessment will force teachers to teach differently, and, in various ways and to various extents, they do. Each assessment effort will have both positive and negative consequences. The design and promulgation of an assessment program is only an approximation of what actually occurs. The operation described in any report is a partial misrepresentation of institutional initiative and measurement integrity. For a reader, it is an opportunity to misperceive what is happening in the schools and the lives of youngsters. We need better descriptions, better evidence, of those consequences of assessment. And partly because we construct nuances of meaning faster than we invent measurements, we need to understand that we will never have a clear enough picture of the consequences of assessment. All findings should be treated as partial and tentative.

Value determination

Not only has there been an increase in the amount of formal educational assessment but assessment has been applied increasingly to influence the well-being of students, schools and systems. The "stakes" have risen. Funding, autonomy and privilege have been attached to levels of scoring. The intention has been to get students and teachers dedicated to their tasks, and this sometimes happens, but there have been costs as well as benefits. Among the reported negative consequences of raising the stakes of assessment are:

a. instruction is diverted,

b. student self-esteem is eroded,

c. teachers are intimidated,

d. the locus of control of education is more centralized,

e. undue stigma is affixed to the school,

f. school people are lured towards falsification of scores,

g. some blame for poor instruction is redirected toward students when it should rest with the profession and the authorities, and

h. the withholding of needed funding for Education appears warranted.

The most obvious consequence of increased assessment is that teachers increase preparation for test taking, including test-taking skills and greater familiarization with the anticipated content of testing. Also, topics tested are considered of higher priority and topics untested slip in priority. Assessments are not diagnostic. There is little strategic theory fitting pedagogy to assessment so that few teachers know how to respond to poor student performance, other than to try harder. Thus, over-emphasis on assessment erodes confidence in legitimate teaching competence.

As the stakes rise, the central authorities are both pressured and authorized to intervene more in teaching responsibilities. A widespread public perception of legislators and school authorities is that they are not knowledgeable or competent in matters of the classroom. With ever-confirming evidence that students continue to be testing poorly, the public is tempted to withhold funds for needed improvement in instruction. There is good evidence that increased funding alone will not greatly change the quality of teaching. But at the same time, by investing in the assessment of students without investing in more direct evaluation of teacher and administrative performance, the professional people and the elected overseers are partly "off the hook." In summary, the consequences of assessment are complex, extending far beyond the redirecting of instruction toward state goals.

It is too much to expect that we soon will clearly discern the consequences of assessment and, even less soon, what caused them. Both the consequences and the causes are complex, both as to constituents and as to conditions. Lacking an adequate research base, curricular policy needs to be based on deliberations, long and studied interpretation of assessment, experience, and ideology. That is unlikely when professional wisdom is getting little respect. Sometimes the public presumes that educators put their own interests above those of students. But good deliberations are not uncommon among school leaders. Evaluation of the consequences of assessment has an important role informing those deliberations.

Even if we were able to improve determination of the consequences of assessment, we lack theory and management systems that guide us in applying that information to the improvement of teaching and learning. We need not wait for politics or the professional to be reformed. We can rely on the political, intuitive, and leadership processes we now have to make assessment more a positive and less a negative force within education.

As indicated before, people do have different purposes for education and for assessment. And for any one purpose, they value the results differently. That is just part of the reality, neither excusing nor facilitating the assessment of assessment.

The assessment practice that does the most measurable, immediate good is not necessarily the practice that has the best long- range effect. For example, using testing time entirely for easily measured skills instead of partly for “ill-defined” interpretive experience increases precision and predictive validity but discourages well-thought-out advocacies to include problem-solving experience throughout elementary school. Value trade-offs need to be considered for long-term as well as short-term effects.

Curriculum and instruction

Management of teaching and the curriculum cannot be effective without assessment. The best and the worst assessments we have are informal and teacher-driven, sometimes capricious and sometimes more aimed at avoiding embarrassment than maximizing services to children. Yet, it works pretty well, sensitive to what individual children are doing, viewed favorably by a substantial proportion of parents and citizens, especially those people who interact themselves, even in small ways, with the academic program. Still, instructional assessment could be much better, and too little professional development is so aimed. The present informal assessment system is little engaged with the formal management information system of school districts and even less with the state’s student achievement testing apparatus.

The most successful school improvement efforts have been those that decentralize and protect authority so that a match can be made between what the teachers want to teach and the parents and immediate community want taught. The recent “national standards movement” was a step in the wrong direction, a further imposition of external values. Assessment was used to nullify decentralization efforts. The state does have a stake in what every child is learning but the state is poorly served by having each child trying to learn the same things. Accountability of the schools is in no way dependent on having each child tied to a core curriculum and tested on the same items. A single test for all is cheaper, but not a service to a diverse population of children. It is not important for good teaching to have access to test score comparisons among students.

State assessment is not wrong in its most general finding that teaching and learning in the American schools are mediocre. And that the range across districts is huge. The spread of achievement scores is stable and predictable, more a function of a child's lifetime educational opportunity than of what happens during a year in a classroom. Neither massive changes at home or in the classroom are likely to result in substantial gains on current assessment instruments.

As stated earlier, the validity of measurement of achievement is not the same as validity of those same scores as an indicator of quality of teaching and learning conditions. Teaching can be changed in a number of important ways within a school or classroom without getting a change in achievement means. Using those scores as a measure of school improvement has not been validated. No accumulation of evidence shows assessment to be an indicator of good schooling. In spite of the absence of validity, comparing assessment statistics continues to be the primary criterion for reform in a vast number of school districts. Given vigorous school improvement efforts over 20-30 years within countless districts, essentially all of them unaccompanied by substantial change in assessment results, what should be concluded is that testing is insensitive to important changes in teaching. Does it mean that schools cannot be improved? No.

Uses and stakes

The uses to which assessment information will be put varies not just across assessment approaches but greatly within approaches as well. Different school systems, teachers, and children, even those greatly alike, will be affected differently. It is not reasonable to suppose that the stakes of assessment are unimportant if they have little impact upon the majority. Special attention needs to be given to how assessment consequences affect the least privileged families and most vulnerable children.

One of the primary stakes of testing is the well-being of teachers. Teachers have much to lose in a high stakes assessment system. Assessment should not be avoided just because teachers protest but their working conditions and professional wisdom should not be trivialized. Teaching quality should be scrutinized. Student performance should be considered but it should not be a primary determinant of teaching competence. There is only a small connection between how well a teacher teaches and how well the class performs on a test.

One of the consequences of high stakes testing is the manipulation of rosters to excuse poor scoring children from participation. The most common way at present appears to be to have children classified as “special education” students. A good bit of ingenuity has been shown in manipulating rosters.

High stakes assessment sometimes does result in raised scores but the validity of widespread gains, locally or across the country, has seldom been established. No one wants to challenge the gains that appear, but presently emphasis on small changes serves to orient the school to the assessments rather than to education. Many of the consequences of assessment are best learned from the people who administer the tests, even though they have self-interest. Many are quick to acknowledge that the assessment enterprise is flawed.

Good research can help but it is mostly a professional and political matter. Until community attitude sets out to make the best of the schools, less to blame them, (however much they deserve the blame), not much good will happen. This is not a nation dedicated to the best possible education system. There are lots of people who would rather have lower taxes than to extend educational benefits. Higher taxes do not assure better opportunities but an interest in finding better opportunities is not a national purpose. Looking at it simplistically, support for assessments appears to be a step toward improving education, but the quarter-century record shows that assessment-driven reform has not worked. Why does it continue to be politically popular? The main consequence of assessment-based reform is that education has not substantially improved. We have plenty of evidence of that.

2000

I wrote about ethics and equity on Martin Luther King Day, 2000.

Dreaming His Dream?

In 1929, my mother took me to the Nebraska State Fair for a physical examination. In those last pre-Crash days, the state had a "Healthy Baby" contest. I won a ribbon, down-pointed only for "open-mouthed, not very intelligent looking." I long supposed it an ordinary competition, like Cornhusker football or Aksarben horse racing, but came to admire the State Department of Health for cleverly providing free examinations for babies of the poor managing to get to Lincoln.

Reading Steven Selden's *Inheriting Shame*,[1] I realized another possible motive. At the Kansas Free Fair in Topeka the same year, the American Eugenics Society displayed racist posters including one that said: "Every 15 seconds, $100.00 of your money goes for the care of persons with bad heredity (p 25)." Had my father's store been several miles south, in Kansas, my examination might have been a step toward future distinction as to who should be encouraged and discouraged -- or disallowed -- from propagating. Maybe it was anyway. Eugenicists such as Carl Brigham claimed, according to Seldon (p 109), that "American ethnic diversity was a threat to national welfare" and cast their hereditarian prescriptions in racial terms.

My mother was proud of her Nantucket ancestors and started me, with charted descendent lines of Tristram Coffin, on a lifetime search up the family tree. We tended to avoid off-

[1] Selden, S., 1999. *Inheriting Shame: The Story of Eugenics and Racism in America.* NY: Teachers College Press. Page 25.

island marriages. I had to go back 25 generations to find someone as odd as a Catholic.

So perhaps it wasn't by accident I was fascinated by psychometrics. I liked winning. I liked wordplay. One way of winning is to invent the game. For my Masters thesis, I developed a quantitative aptitude test, which was used by a small number of graduate schools of education deciding who should be admitted for advanced study. I sat uncomprehending and open-mouthed through most of my psychometrics classes at Princeton, but my professor scored only 95% when he took my QED test. I looked forward to a career of teaching others how to discriminate between the more and less talented.

Brigham was one of the founders of psychometrics. With others, he analyzed the scores of the World War I Alpha intelligence test, used by the Army to decide who, regardless of heredity and wealth, should be sent to officer's training. Those of Northern European heritage as a group did better than those from farther South. I have always felt Nebraskans superior to Kansans.

As with the Scholastic Aptitude Test (SAT) fifty years later, it was widely presumed that if a test predicted which candidates would do well in later classes, it would also identify who would contribute to science, business, government and social service. Brigham and the early psychometricians were proud of their validity studies, although they concentrated on criteria of school performance and not on the battlefields of industry, and nurturance.

Academic success has long been held sacred. One of my mentors pointed with pride at a fellow Nebraskan, Leta Hollingworth of Teachers College Columbia, an early advocate of gifted education. Professor Hollingworth said, "Modern biology has shown that human beings cannot improve the qualities of their species, nor permanently reduce its miseries, by education, philanthropy, surgery or legislation." [2] It was her contention that society should concentrate on schooling the children of able parents, with tests used to identify a few who could rise above their heredities.

The claim was similar to the social engineering argument of James Conant in turning university admissions offices toward the intelligence quotient, later called scholastic aptitude, for selection of students. But, as Nicholas Lemann[3] pointed out in *The Big Test,* those who scored high and availed themselves of scholarships did not, as Conant promised, turn their lives toward public service but used their subsidized education to avail their families of the good life. Another instance of Northern Europeans winning by inventing the game.

I have admired the ingenuity of many teachers of the gifted. And I have been dismayed by the poor engagement of fast learners, and slow learners as well, in many a classroom. Gifted teachers and special education teachers especially have maintained commitments to the individualized learning plan.

Differentiation, tracking, unethical? No. I have at times sided with Brigham and Hollingworth, feeling that those who would vie with the teacher for control of the classroom should be tracked out of it. Oh, not denied an education but separated. Separate but equal.

Did I say that? What do I believe? I believe every child should have educational opportunities based on need, readiness and appetite. Parents and the state should be served too, not by standards and conformity, but by uniqueness and engagement.

There are lots of special groups. At Arizona State's RACE 2000 conference recently, philosopher Ed Gordon told me his son told him, "It's those seven bad dudes that society needs to

[2] Ibid, p.198.

[3] Lemann, N., 1999, *The Big Test.* NY: Farrar, Straus, Giroux.

give its best to." Not because we, like the eugenicists, fear them but because they are our children and because they have in their deviance revealed a talent. Have we no use for nonconformity?

But a mother of twins asks a teacher, "Why did you teach Sammy how a camera works and not Sally?" The teacher assures that Sally needed to work on her math. And the mother says, "It isn't fair for you to give your better self to one rather than another."

Every effort to customize, to deal with the uniqueness of a human being, results in discrimination. We know no way to teach separately and equal. Yet we know no way to do right and treat all the same. Together we must be at our most ingenious selves to devise the meaning of equality, partly because it must mean special blessings for those who haven't invented the game. Neither tests, nor heredities, nor unfinished assignments justify privileging one child over another, one family over another, one race over another. Would that we would honor diversity and equal rights, and had the intelligence to write a curriculum worthy of our dreams.

2000

This was my brief panel presentation at the annual meeting of the American Evaluation Association, 2000. To illustrate an interpretive approach to program evaluation, I used a vignette from a case study of Project Heart operating in the Champaign schools. The Getty Trust had sponsored a small collection of case studies illustrating the teaching of elementary school art along the lines of its Discipline Based Arts Education.

Had I waited a few more years to speak, I would have included these evaluation books: my Standards-Based and Responsive Evaluation,[1] *edited and promoted by Deborah Laughton and published by Sage, Saville Kushner's* Personalistic Evaluation,[2] *and Merel Visse's* Evaluation for a Caring Society.[3]

Evaluation Research Strategies

The unique thing about evaluative research is that it puts the spotlight on quality. The principal questions are questions of merit and shortcoming regarding some evaluand, usually an evaluand at a particular time and in a particular context. Some evaluators focus on impact, or effectiveness, or productivity, or costs and benefits -- but they all treat those things as criteria of goodness.

All kinds of inquiry methods are used in evaluation studies and the good studies will use more than one method. Qualitative evaluators like myself focus more on interpreting

[1] Stake, R. E., 2004. *Standards-Based and Responsive Evaluation*, Sage,

[2] Kushner, S., 2000. Personalizing evaluation. London: SAGE.

[3] Visse, M. & Abma, T. (editors), 2018. *Evaluation for a caring society.* Information Age Publishing.

what it is that is happening, that is, on the process. Many quantitative evaluators focus on outcomes, using measuring instruments of some kind to get data on the effects of the program.

To study the quality of the development of something in order to improve that development, we use formative evaluation. To study something already developed to know its value for use, we use summative evaluation. “When the chef tastes the soup, it's formative. When the guest tastes the soup, it's summative evaluation.” These may be my most quoted words.

Educational researchers do evaluation in lots of different ways. Some call their methods "models" of evaluation but, because I think them too loose to be called models, I call them "approaches" or "persuasions" or “predilections.” I think the most important distinction is between the analytic and the interpretive approaches. Usually the interpretive or qualitative evaluator immerses self and team in the workings of the program, maximizing experience with it, pursuing complex questions, some of which don't have an answer, then prepares a discussion of the quality perceived. Usually, the analytic evaluator identifies a few criteria, notes needs and standards, obtains instruments or protocols, objectively gathers large batches of data, and submits them to statistical analysis. The findings of one are more likely to be precise and superficial; the findings of the other are more likely to be subjective and profound and irritating. In both approaches, the better evaluators will find ways to repeat their observations and challenge their biases.

A few of the best references and resources (as of 2000), I would say:

The American Evaluation Association Discussion List <JOV ALTALK@BAMA.UA.EDU>

The American Journal of Evaluation, published thrice a year, the voice of the American Evaluation Association.

Evaluation Models, Second Edition by Dan Stufflebeam, George Madaus, and Tom Kellaghan, published by Kluwer.

The Evaluation Thesaurus, Fourth Edition, by Michael Scriven, published by Sage.

Foundations of Program Evaluation, by Will Shadish, Tom Cook, and Laura Leviton, published by Sage.

Professional Evaluation, by Ernest House, published by Sage.

As to typical data that I regularly use in my interpretive approach, here is an episode illustrating an issue of importance, this one regarding how Larry Ecker, a visual arts teacher in the classroom, is directing student attention to production and planning. It opened my part[4] of the 1984 report to the GettyTrust.

Noisily, in twos and threes, the sixth graders cascade down the stairs and into the basement room. "Okay, have a seat." All sit except Bonnie who for some reason closely examines the pencil sharpener. "Attention right here please," says Mr. Ecker. "Remember yesterday we made drawings from three sources: fantasy, observation, and recall." One child says, "Flashback." "Yeah, like flashback." (pause) "Mark, you made a promise. Okay? Vince, put your foot down. Listen. I will not let one or two of you spoil it for the bunch." Heads bow just a bit.

"Now, get your drawing of yesterday in mind. Which category did you use?" He talks with them about what they drew, why they drew it. All eyes all him. "What made your

[4] Stake, R. E., 1984. An Illinois pair: A case study of school art in Champaign and Decatur. In M. Day, E. Eisner, R.E. Stake, B. Wilson, and M. Wilson (Eds.), *Art history, art criticism, and art production*. Santa Monica: Rand Corporation.

drawing look good? You changed your mind, didn't you, as you went along. What told you that something should be changed?"

"Today I want you to start on a new drawing. We'll finish it tomorrow, or maybe not until Thursday. I want this one really good. I want you to use a combination of observation, fantasy, and recall, all three. It must be a full-page drawing, touching all four sides of the paper. You should start in pencil, then finish it in color."

"For the observation I want you to draw the shoes you have on today. One or both shoes. As long as your socks are reasonably clean, you may take them off." Responses of "Phew!" and rolling eyes.

"Now for the environment, something out of fantasy. Maybe your shoe could be swimming in an aquarium. Or it could be in a space suit on the moon." Several questions "Could it be ... ?" "Absolutely. The more outrageous, the better. Okay, get your materials and do it."

The description and insightfulness of Ecker's teaching continued for pages.

2001

I submitted this piece to the NY Times *for the Op Ed page. It wasn't published.*

"That'll Learn 'em!"

Reason is taking a caning in the management of American schools. Congress is taking its own steps toward disciplining American teachers. According to the *New York Times*,[1] New York State Commissioner "Rick" Mills has punished 34 New York City alternative schools for failing to use Regents Exams to decide what and how they will teach. The 34 had been working together, at times led by the Urban Academy's Co-Director, Ann Cook. In a PBS interview, Ann said:[2]

I think what you have to look at is not only whether students can score well on a reading test, but whether they do read. Do they use that skill? Is it something that they enjoy? Would they go and get a book in their spare time? Now, if you have students who start off and who aren't reading, what happened to them, when they graduate has that shifted, has that changed? Has the school done anything to get kids more interested in reading?

Mills and the public have reason to be disappointed with the schools. Many are mediocre; the variability is large. Many are prisons of mind and soul, insisting on conformity and standardization. Many have rebels who vie with authorities for control of the classroom.

[1] Holloway, L., *New York Times*, April 26, 2001.

[2] https://www.pbs.org/onlyateacher/today1.html#top

There is reason for reform, but using tests that only poorly measure good teaching and learning has not worked. Reform by testing has been tried unsuccessfully in Michigan, Florida, California and in many states for a quarter of a century. It has not worked. It is unreasonable, and is forcing education further into frailty.

I was a member of a so-called Blue Ribbon Panel advising Commissioner Mills on the alternative school's use of their performance testing in lieu of the state-mandated Regents examination. He chose to disregard our advice and moved to prevent those and other schools from using a form of teaching that was working well with a variety of youngsters, and as research showed, lowering the drop-out rate and getting them into colleges well above their-own-zone schools.

It is the view of many politicians, marketers, editors, parents and more than a few teachers that standardization of goals and lessons is a fundamental virtue. But smaller schools, project-teaching selectively tied to curricular standards, portfolio-making, and individualized performance testing contribute much more than standardization to education and maturation.

Our report did not advocate individualization of instruction. We recognized indications of success of the schools using individualized performance assessment and urged further study. We reported that we did not have good data for evaluating the merits of the schools own assessment, but could not ignore the testimony of students, teachers, principals and parents involved.

Our panel members were particularly impressed with the claim of the principal of the Brooklyn International School that children with weak English needed to spend great amounts of time speaking English, even at the expense of coverage of the state's content standards. We did not apply the same reasoning to children who have weak thinking skills. But should we not have?

Forms of critical thinking are as important in third grade and eleventh grade as they are in graduate school. In these schools, performance assessment emphasizes problem-solving. Most Regents schools do not. Teachers in the 34 alternative schools coach the students on academic projects, seeing their recognition of problems, judging their choices, linking them to other curricular goals. They convince most of their advisees that this is the essence of learning and on what they should be graded and graduated. As they see it, a "high stakes" test requires days of preparation and undercuts the motivation of students to stay in school and be autonomous learners.

The New York Regents tests, the currently mandated tests of many states, and those proposed by President Clinton are not measures of what a student has achieved but of scholastic aptitude, a certain general readiness to learn. Even with questions worded as history or mathematics questions, figuring them out is more a matter of general aptitude than subject matter mastery. Such aptitude does change some over time but it is little indicative of a semester's good teaching. Current tests are not a valid measure of teachers or schools, and certainly, "high stakes" application of the scores is a violation of reason. Our report noted the lack of research to justify graduation decisions made from Regents scores.

It isn't only ignorant parents, administrators, and teachers who, in technical language or crude, mumble "that'll learn em!" as they punish children for not learning. New York City's 34 alternative schools have demonstrated ways of getting students to want to learn, not needing Regents scores to show when they are ready to graduate.

New York and the nation have taken a political stand, supposing that performance on standardized tests is a guide to making schools better. Those tests argue that the students collectively do not have the right aptitudes for schooling. They do reinforce the view that we need better schools. But presuming then that preparation for testing is the best curriculum is a political option, a punitive option, a hurtful option, supported neither by professional experience nor educational research.

2002

I've given lots of thought to family trees. More to names and dates than personhood. Of the maybe 2000 names I now have on my tree, almost all with birth and death dates, all from a home-place east of Seattle, west of Rome, north of Gibraltar and south of Trondheim. Quite a few were Crusaders or into other miscreancy. A Grandpa taught Indians, a few Christian missionaries in the Far East. We were proud of them, until we learned about feudalism and colonialism and the Westward movement. It now seems possible to become shamed in the eyes of my own grandkids.

Conception

Nellie is my four-year-old granddaughter. She calls me Grampy. When we race in the park, she beats me fair and square.

Granddaughter is a concept. I have had other granddaughters. Ah, yes, every woman has been a granddaughter. So when I think of granddaughters, I not only think of Nellie, but Laura and Alison and, to a small degree, every woman who has ever lived. They all spring from my mother lode of granddaughters.

Nellie is real and specific, so technically speaking, she is not a concept. But my thoughts of Nellie go beyond her home-place and age and gender and red hair. I associate many things with Nellie. When she says she wants to be on my team, it warms my heart, even as it cools when I realize she prefers Grammy's lap.

So Nellie is a concept too, something beyond denotation. The concept of Nellie includes the connotations.

There are two generations between us, two mothers in their thirties, mothers in some families old enough to be grandmothers. You don't put those things on genealogical charts.

It is difficult to measure how smart my ancestors were. My 70-year old brother claims he is smarter than Einstein -- staying alive, as he is, longer. Smart yes, but as educated as Nellie? I can't tell. It's a concept needing work.

2003

This is a statement on advocacy, activism, confluence of interest, and uncertainty, perhaps with a surprise ending, a paper I delivered at the Annual Meeting of the American Evaluation Association, Reno, Nevada, November 6, 2003.

How Far Dare a Program Evaluator Go Toward Saving the World?

No two professional evaluators are the same but many use similar methods. Still, each of us will use a method in a somewhat idiosyncratic way. Especially in the interpretation of data, personality and experience have a play.

Professional evaluators come from many backgrounds. They have greatly different aspirations. As a group they are considerate people. They are ethical. They follow disciplined procedures to find the merit and worth of a program or other object. Oh, there is a rogue here and there. He or she may go where the money is. Or the next contract. But most of us evaluators are good people, most of the time. We are specialists at recognizing differences among greater and less quality. We hope that our work contributes to the making of a better world.

While sitting in a waiting room, I overheard a young woman say, "I'm an evaluator now." Answered by "Don't you do consulting any more?" "Oh, yes, but I get more attention if I call it 'evaluation.'" Some evaluators call their work "evaluation," thinking it gets more opportunity to help make changes for the better. In large and small ways, they hope to help save the world. Is this false advertising?

One evaluator I know is passionate about discovering perfidy, particularly bureaucratic deceit. Another I know seeks to balance evaluator voices with stakeholder voices. As for me, I find myself digging into issues of continuing professional education, regardless of the questions prioritized in the contract.

So I speak of *advocacies.* Most evaluators claim to make dispassionate searches for quality and dysfunction. They speak disdainfully of advocacy and self-promotion. Yet it is clear that most of us evaluators have strong feelings about certain matters which we favor in our work[1]. Here are six advocacies common in evaluation studies:

We care about the evaluand, the object being evaluated. Often we believe in it. The *internal* evaluator is evaluating a part of his or her own organization. Barry MacDonald once told me,[2] "One should not evaluate a program if one does not support its goals." Occasionally we have a conflict of interest; more often a *confluence* of interest. We *hope* to find the program working. Most of us are *disposed* to see evidence of success more quickly than evidence of failure.

We care about evaluation. We want to see others care about it. We want to encourage them to do it. We promote evaluation services, our own and those of our profession. We favor methods that evaluate well, and encourage others to use them too. It is an advocacy we flaunt.

We advocate rationality. We would like our clients and other stakeholders, our colleagues and heads of department to explicate and be logical and even-handed. We often pause in our data gathering or reporting to point out a way that the evaluand could have been run more rationally.

We care to be heard. We are troubled if our studies are not used. We feel evaluation is more useful if program participants take some ownership of the evaluation. Many of

[1] Mabry, L., 1995. Advocacy in evaluation: Inescapable or Intolerable. AERA annual meeting, Vancouver, BC.

[2] Personal communication.

us, including myself, are strong advocates of self-study and action research. Even an external evaluation can profitably use input from stakeholders -- including suggestions for design and interpretation. Many of us, not including myself, strongly support participatory evaluation in which certain stakeholders take responsibility for design, data gathering and resolving questions of merit and shortcoming.

We are distressed by under-privilege. We see gaps among privileged patrons and managers and staff and under-privileged participants and communities. We aim some of the evaluation at studying issues of privilege, conceptualizing issues that might illuminate or alleviate under-privilege, and assuring distribution of findings to those often excluded.

We are advocates of a democratic society. We see democracies depending on the exchange of good information, which our studies can provide. But also, we see democracies needing the exercise of public expression, dialogue, and collective action. Most evaluators try to create reports that stimulate action.

These six advocacies are easy to find in evaluation reports. Although we are troubled by the possibility that our advocacies will cause us to search more vigorously for aspiration-based evidence than other, we cling to some advocacies more than to neutrality, believing these well-considered biases to be compatible with the interests of the profession, our clients, and society.

Each of us is more than an evaluator. We are complex human beings. Some of the things we do are part of our work and some are outside our work. We have political, spiritual, aesthetic and other advocacies. Some of the panorama of advocacy cannot help but become part of the evaluation study, even if we try to confine it to the other parts of our life. Perceptions and values from any part of our lives may influence the interpretations we make.

Ethical standards

It is an ethical responsibility for the evaluator to identify the affiliations and ideological commitments that might influence his or her interpretations, not only for the contractors but for the readers of reports, and of course, for the evaluator her- or himself. But there is no way for the evaluator to identify all predispositions, nor even *to know* them. We can expose ourselves a little, through vitas, biographical notes, previous reports, acknowledgements of preference and alliance, even indirectly in the ways we write, but an entire list of possible influences would be arbitrary, ever out-of-date, and ever incomplete. It is difficult to help others realize more than a few of the biases, good and bad, to be found in our work.

We evaluators sometimes see the conflict in two main roles: judgment and remediation, one the role of evaluator as finder and reporter of program quality, the other the preservation of quality and restorations to quality. A contract to discern quality *is not* a license to fix things. The client may want help fixing things, and may so specify in the contract, but there is reason to refuse. Evaluators as program improvers are under some persuasion to over-attend to things they can fix and to neglect things they cannot. They may apply for the evaluation work mostly to get the remediation assignment. They may do good things for the program but leave it not thoroughly evaluated. The client may claim the program was evaluated when what really happened was some fixing. There are reasons for keeping the two tasks separate, evaluation and repair.

Generally we presume that successfully carrying out an evaluation study is a contribution to social well-being. Our report will describe program activity and problems and show, to some extent, the merit and shortcoming, its effectiveness, productivity, and impact. Some of that information should be useful to program managers, the staff, the service recipients and stakeholders in general and should add to the efforts of professionals, and reformers and others, to make this a better world.

Sometimes the evaluator is encouraged to instruct the staff in the processes of evaluation. Sometimes the evaluator chooses to supervise the staff's gathering data. Sometimes the evaluator is persuaded by his or her discipline, fellow evaluators, or conscience to help the organization or all society to become a more rational or more democratic enterprise.

Evaluation addenda such as these are often little explored during negotiation of a contract. Some of them grow spontaneously during the evaluation work. Spontaneity and emerging sensitivity can be good but they sometimes violate evaluation ethics calling for full disclosure of purpose and practice. How far should an evaluator go toward saving the world?

I have studied documents of ethics and standards, looking for guidance on matters of advocacy and activism. Are both conflict and confluence of interest recognized as a problem? What do the standards say about advocacies? The best statements of ethics for evaluators are two documents, the *Joint Committee Standards* (1994)[3] and the *American Evaluation Association Guiding Principles* (www.eval.org)[4].

Let's look at those *Guiding Principles.* For my search for support or restraint on evaluator advocacy, I enlisted some brief help from Will Shadish, one of the authors of the *Guiding Principles*[5] He selected as relevant the two principles I cite

[3] Joint Committee on Standards for Educational Evaluation, 1994. *The program evaluation standards. How to assess evaluations of educational programs.* Sage.
[4] Greene, J. C., 1996. Qualitative evaluation and scientific citizenship: Reflections and refractions. *Evaluation, 2, 277-289.*
[5] Shadish, W. R., Newman, D.L., Scheirer, M. A., & Wye, C. (editors) 1995. *Guiding Principles for Evaluators.* New Directions for Program Evaluation, 66. Jossey Bass.

below, plus adding that Principles C1, C4, E3, and E5 suggest that "an evaluator should inform a client if there is reason to believe he or she might object to a particular value commitment." The most relevant two were:

A2. Evaluators should explore with the client the shortcomings and strengths of both the various evaluation questions it might be productive to ask and the various approaches that might be used in answering these questions.

Hmm. Nothing much there.

C3. *Evaluators should seek to determine, and where appropriate be explicit about, their own, their clients' and other stakeholders' interests concerning the conduct and outcomes of an evaluation.*

Well, that's a start. No constraint or support, just acknowledgement. The *Principles* do not directly show concern about evaluator advocacy, nothing drawing attention to the six biases I mentioned earlier. Shadish pointed out that the drafters of the *Principles* were careful not to deal at the level of particular methodological approach, seeking broad positions at a higher level of conception.

The *Joint Standards* were equally non-specific about evaluator value-orientations. Standard C3 calls for full and frank disclosure, C4 calls for balanced reporting, and D11 calls for objective reporting (pages 74, 77, 138z). The document goes on to identify guidelines and pitfalls, getting more specific, but still not raising explicit concern about confluence of interest, intent to remediate, or activism. There is *implication* that one should stick to the job of finding merit.

Now let us expand the concern about guidance regarding advocacy of personal values to advocacy contained in methodological approaches. Let us look briefly at two currently attractive values-committed approaches: participatory and democratic evaluation. Participatory evaluation (in independent external evaluation) is aimed at engaging the program staff in responsibilities normally belonging to an evaluation specialist, particularly: research design, data gathering, analysis and interpretation. [6] [7] [8] It is claimed that involvement of the staff will make the study more relevant and increase the likelihood that the findings will be used for further program development. Sometimes program beneficiaries and other stakeholders also are invited to participate. Many of us expect that the quality of such an evaluation will be technically inferior and, in some ways, conceptually less deep. But the organization may be helped much more by participatory evaluation than by a conventional external evaluation. Advocating stakeholder power, the participatory supporter claims that improving the organization can be worth more than broad determination of merit and worth.

Democratic evaluation is built on another powerful advocacy of some evaluation writers. [9] [10] [11] This approach draws the attention of the evaluator to concerns of under-represented stakeholders, the people seldom heard, sometimes seeking ways of engaging them in dialogue about policy and program change. The evaluator may merely assure that some issues reflecting the values of the under-represented are included in

[6] Patton, M. Q., 1996. *Utilization-focused evaluation*. Sage.

[7] Kemmis, S., 1986. *The action research planner.* Geelong: Deakin University Press.

[8] Fetterman, D. M., 1994. Empowerment evaluation. *Evaluation Practice, 15,* 1, 1-15.

[9] MacDonald, B., 1977. A political classification of evaluation studies. In Hamilton, D., Jenkins, D., King, C, MacDonald, B., and Parlett, M., *Beyond the numbers game.* London: MacMillan.

[10] House, E. R. and Howe, K. R., 1999. *Values in education and social research.* Sage.

[11] Greene, J. C., 1996. Qualitative evaluation and scientific citizenship: Reflections and refractions. *Evaluation, 2, 277-289.*

the design or may, in the interpretation of findings, strive to make a persuasive case for their support. And he or she may try to make findings available especially to people who might contribute to that support. Democratic evaluation is clearly on the track of trying to make a better world.

Can diversity be a standard?

Efforts to improve the evaluand and the world will diminish attention to the primary evaluation issues. Conceivably the evaluator could confine remediation and social enhancement to off-duty time, to that other life, thus outside the formal study, but that's not the way it's done. Whether in democratic evaluation or the exercise of any other advocacy, the evaluator makes it part and parcel of the evaluation work. And it is impossible for the evaluator's interpretation of findings not to be colored by any parts of his or her life and camaraderie, community, and culture.

The Joint Committee Program Evaluation Standard A5 (page 37) says:

> *The evaluation report should describe the object being evaluated and its context, and the purposes, procedures, and findings of the evaluation, so that the audiences will readily understand what was done, why it was done, what information was obtained, what conclusions were drawn, and what recommendations were made.*

This is noble aspiration but the evaluator is not able to tell the full story. The reality is that, exercising all good faith, much will not be reported, and some will not be understood.

Again I say, much will differ from evaluator to evaluator. Most of us aspire to a professional practice by which – hypothetically -- all evaluators evaluating a single evaluand would produce largely the same findings. But it is not an attainable aspiration, and to force it to happen would invite disaster. Evaluators cannot help but see some things differently. Some findings will be different. Hopefully not often completely at odds, but that too will happen. In the complex determination of program quality and accomplishment, there is no single reality we can capture. Reality is constructed by people, and people sometimes differ. When we agree on what we see, we tend to think we see correctly, but sometimes we do not. When we disagree on what we see, we tend to think one of us sees incorrectly, but sometimes both see correctly.

We have an evaluation practice that is influenced by the value commitments of the evaluator and a set of operating standards that imply we can attain a widely-agreed-upon picture of merit and worth. Something has to give. It could be that we should more effectively constrain our value commitments and search harder for meta-evaluation consensus, but we clearly should develop our standards and principles so that they deal better with the uncertainty and individuality of evaluating.

One of the guiding principles should say something like:

It should be expected that any two competent evaluators, working together or apart, will seldom agree fully on criteria and standards, critical incidents and experience, and on the appropriateness of the evidence of merit and worth. The full use of validation, triangulation and meta-evaluation is essential but it will not eliminate disparity in the evaluation findings.

Evaluators should be encouraged “to have a life” and “to have a dream,” so their interpretations are enriched by personal experience. Idiosyncratic interpretations are small steps toward saving the world.

2006

As I review manuscripts or read student reports, I try to consider each as a totality, and as an individuality. Only secondarily do I try to look at them analytically, satisfying this or that criterion. It is the totality that gets priority. What kind of creature is it? I try to realize the special uniqueness of each paper, realizing how it might be different if done at a different time or by a different writer. At least at first I resist the idea that the quality of a case study is some aggregate of scale values. I look for the quality of the whole. But as I look back over my reviewer remarks, it is apparent I am analytic too. I do have criteria. I note certain "short-comings" occurring over and over. The derivative criteria are listed below. Any one of them might be unimportant for a particular case study. But again, any one might almost completely determine a study's lack of quality.

Criteria for Naturalistic Case Studies

Conceptual Structure. When I have finished reading, I want to feel that I know the structure of ideas, the theoretical framework or emic issues, around which the report was organized. Mere chronological or institutional or goal structure usually is not sufficient. Still, I do not want to feel that the conceptual structure is so strong that it has kept key observations from being made.

Editorial Organization. I want to feel that each paragraph has been reread critically, finally assembled in a way that contributes to the whole. The study should have a sense of closure, probably ending on the ideas of its opening; with good format.

Author's Viewpoint. I prefer to be told, one way or another, the author's experience and feelings. If these do not nicely fit into the report, then footnotes, a preface or even a cover letter can be used. I often am dismayed, however, with extensive subjective ruminations or with an author's frequent references to self.

Sense of Audience, Place and Context. I like to feel that I know to whom the author is writing. I like to feel that I know where the action is happening. I like to feel that I know the social, historical, political, educational, economic conditions.

Over-interpretation. Naturalistic case studies are interpretive research. Still, I want raw observations available for the reader's interpreting. To say, "The father was indignant" or "It was an efficient office" is not an observation, it is an interpretation. Facts should come first. Over interpretation is one of the most common and distressing failings of case study writing. Also, overgeneralization is bad, such as concluding that “teachers do” rather than “these teachers did.”

Failure to Interpret. After the observations are presented, perhaps at the very close of the report, the author should give additional interpretation of things according to his or her special purview or competence. For example, an educator should acknowledge relevant educational writings. (Passing judgment or indicating **causality are seldom necessary interpretations.)**

Validation. I want the author to let me know, directly or indirectly, how the observations were validated, and even what efforts were made to disconfirm principal conclusions, to tell what deliberate efforts there were to triangulate.

Multiple Realities. I hope the study acknowledges alternative ways of seeing.

Proportion. I want to find a sense of proportion, that the author is in touch with how people value things. To do this requires a good balance among descriptions of ordinary events, issues, contexts, and assertions. The trivial can be reported but not idolized. The momentous should not be ignored.

2009

Arranged by Professor Maria Bustelo, my Madrid presentation here was to a gathering of members of the Department of Evaluation (Instituto Nacional de Administración Pública) and students of the Complutense University in Madrid on April 25, 2007.

Criterial versus Experiential Evaluation

I have many stories to tell. And aren't stories friendlier than specifications? But I am going to use this conversation to make a "hemispheric" distinction between *criterial* and *experiential* evaluation. Hemispheric is more than an ocean apart. I mean worlds apart. Criterial, experiential -- worlds apart. Then a story.

Formal evaluation

Some of us, you over there and you over there, today or in the future, will call ourselves, "professional educational evaluators." We are people who make formal studies of the quality of educational programs. We all are trying to learn to evaluate educational programs better.

Teachers, program administrators and many others also evaluate those programs. Most of the time, they do it informally. We professional evaluators boast, "Sometimes *we* can see the quality more clearly, or find it in different forms, or find it more reliably."

But we know too, that people with special experience -- people such as teachers and department heads and care-givers of all kinds, even our children -- can evaluate some things better than people with formal evaluation training. Fortunate is the program evaluator who knows how to use the assistance of people with special experience!

People ask many evaluation questions: How good is the teaching? How safe are the play spaces? How honest is the report? How long will the milk stay fresh? Was that a good experience for the children? People evaluate all day and every day. We professional evaluators look for *better* ways to discern quality, and better ways to describe to others the quality we find. We look for ways to persuade the readers of our reports that our scores are pertinent and our interpretations trustworthy.

Around the world, today as in years past, much formal evaluation has a strong political side. Many people seek evaluation findings to promote their causes. Many people, including sponsors and agencies, do what they can to configure the evaluation design so that it will *support* their policies. The world of professional evaluation is infused with politics. Does that mean that evaluation reports cannot be trusted? Sometimes.

Criterial Evaluation

Let's examine two ways professional evaluators think about program quality. We have theories and models and practices.[1] After 40 years of working with educational evaluators, I decided that the major distinction in evaluation approaches is between *criterial* and *experiential* evaluation. Not experimental. Experiential. Criterial and experiential.

Criterial evaluation is scalar, quantitative, measurement-oriented, and standards-based. Experiential evaluation is episodic, situated, interpretive, and qualitative. For years I have been calling experiential evaluation, "responsive evaluation." But responsive evaluation has been too much linked to me personally. Today I focus on *your* thinking about evaluation, whether you are evaluating education, or cell phones, or paella. Sometimes you evaluate with emphasis on criteria and sometimes you evaluate with emphasis on experience. Both are good. Both are necessary. Both can be improved.

Criterial evaluation calls for being explicit about the variables, the measurements, the sampling, and the cut-off standards to be used to make assertions about program quality. It emphasizes formal terminology and explication. Often, criterial evaluation focuses on only a few criteria of successful performance, criteria such as student performance on standardized tasks, text comprehensibility, parent participation -- mostly measured simply, perhaps most often called, "outcomes." Criterial evaluation relies on *indicators* of performance. Measurements. Scalar data. We hear criterialists say, "The proof of the pudding is in the eating." It's outcomes that count," so they say.

There are lots of different ways of doing criterial evaluation, most of them quantitative. For example, the school assessment system may choose a single test of student achievement to represent the quality of all aspects of the curriculum. Evaluators do that knowing from research and

[1] Stufflebeam, D.L. & Shinkfield, A. J., 1985. *Systematic evaluation: A self-instructional guide to theory and practice.* Boston: Kluwer-Nijhoff Publishing.

experience that there often is a positive correlation among different manifestations of program quality. So if you measure one criterion well, it will tell you about how well students would do on some other criteria. That is the way the criterial evaluator often thinks: "It's better to measure a little well, than a lot poorly."

In advance, many criterial evaluators try to set standards. How high will the performance have to be for the program to be considered successful? Or taking today's performance, they may compare it to performance at an earlier time, or to the performance of a control group. Sometimes they leave it to experts to decide, after the fact, how good the program performance has been -- but many of them dislike such subjectivity. At least some criterial evaluators are happier when they can be explicit in advance about the objective standards to be applied to decide whether or not the program is good.

Measurements

What many criterial evaluators are most proud of are their measurements. They like to get standardized numbers down on paper to show the performances of participants and beneficiaries. They analyze the numbers, sometimes in complex statistical ways, to show the success of the program. They might show, for example, that -- after adjustments for differences in prior standing and amount of assistance provided -- the students made statistically significant gains. Sometimes that will be seen as grounds for concluding that the program was successful.

But it takes more than that to conclude, with some certainty, that *that* program approach is successful generally, or that future policy should be changed. For generalization, we need to study *variations*_of the approach in a variety of situations. That approach is pretty much the way of ordinary social science and policy evaluation.

I started my career doing criterial evaluation. When I first did a formal evaluation study in 1964, I was a psychometrician and instructional researcher, and I only did criterial evaluation.

But I couldn't make that approach work well enough, so over the next forty years, I changed to being more of an ethnographer and case researcher. And I called this more experiential work, "responsive evaluation." But just for today I am calling it "experiential evaluation."

Experiential evaluation

Experiential evaluation is using personal judgment as the main basis for assertions of program quality. Because personal judgment needs to be based partly on personal experience, experiential evaluation places heavy reliance on examining the personal experience of people, in one way or another, participating in the program -- teacher experience, student experience, the experience of others, including the experience of the evaluator as an observer. When possible, experiential evaluators work *face to face* with the activity, with the problems, with the expectations and ambiguities and contradictions, of the program. Immersed in them.

Usually, quality is best discerned, I say, through experience. Like quality, experience is everywhere. When your mother and father had you as a baby, they made a great contribution to the grand totality of experience. All your life experience is being added to the history of humankind. The fact that other people in the program have different experiences does not make your experience less important. All of them count. And the mean is seldom more informative than the variety. In experiential evaluation, standards are important even when they remain un-spoken. Usually the standards are set intuitively and separately for different people. These standards are based on past and current experiences of the people involved in that teaching and learning.

Yes, experiential evaluation is relativistic evaluation. It is situated evaluation. It is common in daily life, in corporate life, in government life, especially regarding the most important matters.

Formal evaluation, particularly criterial evaluation, pushes the emphasis away from personal experience toward standardized measurement and toward generalizable knowledge. Experiential evaluation works to re-establish an orientation to the experience of individual persons, however small or large the group.

Of course we sometimes go too far in individualizing or localizing the evaluation. Community values are to be taken into account. The values of the people collectively, as expressed in state documents should be important in helping determine the quality of a program. Experiential evaluation is not a commitment only to the values of the individual person but a commitment that the values of the *individual person* are well considered.

Experiential evaluators seek *multiple realities*, the different meanings that different people give to the teaching and learning. They usually feel that one reality is more pertinent or useful than others, but they try to present more than one reality to the readers of their reports. The experiential evaluator usually does not seek a simple summary statement of program quality, but a collection of judgments of quality.

Quick. Think now of early childhood education in Ukraine. Lviv, Ukraine. Svitlana Efimova and Natalia Sofiy,[2] the

[2] Efimova, S.& Sofiy, N., 2004. Inclusive education: The Step by Step Program influencing children, teachers, parents and state policies in Ukraine. Budapest: Open Society Institute.

Ukrainian case researchers, selected seven-year-old Liubchyk as their case, a boy with autism enrolled in a regular first grade. After observing him in his classroom, they stretched far and wide with their observations and interviews to connect with other Step-by-Step[3] activities in Ukraine and other Step-by-Step sites. They interviewed people in teacher training and at the Ministry of Education.

Drawing from her experience in Liubchyk's room, I wrote and gave to her for expansion and translation:

I was visiting the first-grade room of the "Children of the Sun." Together, the children had chosen this title. They liked to say that they were "Children of the Sun." On the classroom door were this title and individual photos of all the children.

It is 11 a.m. Liubchyk is just coming in with his Mom. She helps him take off his coat. Liubchyk is a slender boy, a tall boy, with fair hair and grey eyes. He is eight, a child with special needs. He first started preschool here in Maliuk School in 2000.

Liubchyk goes immediately to the Reading Center. He stays maybe three seconds, then comes to the teacher's table. Ms. Halyna, the teaching assistant, comes across and greets him, "Good morning, Liubchyk!"

He cheerfully replies, "Halyna, at seven!" (which seems to mean that Ms. Halyna should remember to return home from work at seven). He takes several photos of classmates from the teacher's table and starts looking through them. Pointing to a picture, Halyna asks: "Who's that?" "Adij," Liubchyk answers. He starts saying the names of all the persons in the pictures. Then he puts the pictures back in the envelope and returns them to their place on the table. "Halyna, lunch toowelve," he says, pointing to the clock. "Yes, lunch is at twelve," Halyna answers.

[3] Klaus, S., & Ghent, L., 2014. First Steps: A Brief History of the Step by Step Program. Budapest: Open Society Foundations: Voices.

The classroom teacher, Ms. Oksana, is working on mathematics with the group as a whole. The children saw Liubchyk come in, but were not distracted from their tasks. After Oksana gives the children small individual tasks, she approaches Liubchyk to greet him, "Good morning, Liubchyk." "Oksana, at seven!" he replies. "Please say, 'Good morning.'" He does. "Liubchyk, will you work here with us?" He says, "No" and goes to the Reading Center and starts turning the pages of the mathematics textbook. The group lesson goes on.

Mainstreaming children with disabilities is a problematic situation, an extra burden for the teacher. But this report showed the three teachers there easily managing the situation and the children learning extra caring skills because Liubchyk was their classmate. And the case study of Liubchyk helped readers understand how Ms. Oksana changed from being opposed to accepting Liubchyk into her classroom to becoming an advocate for such inclusion.

Comments

And last, a few general comments. The usual purpose of evaluation is not to reach general social science understandings but understandings about a particular evaluand. By understanding better the complexity of the evaluand, we should be better at setting policy and practice.

We should look both for the general and the particular, but each of those aims wants to eat up all the budget. Also:

... Good instruments are very expensive to develop.
... Good observations and interviews take lots of time.
... The things we want most, leave little time for the rest.

Some people will say that collecting “experiences” is not real research and cannot help science. They are wrong. How is fostering professional insight different from building a science of education? Experiential evaluation can help a teacher or specialist reconsider -- during action -- what needs to be paid attention to. New experience changes intuition. Formal knowledge can do the same, sometimes better. Professionals need both, reason and intuition, criterial thinking and experiential thinking.

One of the epistemological strengths of experiential evaluation is the belief that the meanings of activity, such as teaching and learning, adding on and course correcting, are situational. What the teacher is doing is influenced by culture, the home environments of the children, the conditions of the classroom, and the personality of the teacher.

Experiential evaluators sometimes use case studies to probe the meanings of situations and to report to readers the complexity of teacher and student performance. Some of us try to extend to readers a vicarious experience of the program, thus a better opportunity to decide the quality of the program in their own way.

In experiential evaluation there is need for participants and outsiders to interpret what is going on. So the evaluator presents vignettes, pictures, dialogues for discussion, verification, interpretation, seeking alternative meanings. What first appears as a subjective account of happenings -- when triangulated and reasoned through -- can become a trusted part of the report.

I have been talking about what all of you do every day, evaluating things, criterially and experientially. Doing formal evaluation requires both. And each can be done with sensitivity and discipline.

2009

On March 26, 2009, I was invested with the honoris causa degree at the University of Valladolid in Spain. I made the following presentation, speaking of criticism, evaluation, advocacy, human perspectives, honorary doctorates, the Bologna Process, protecting human rights, and the character of a university.

Joining in the celebration were the Rector and his choir, other locals, and, Maria Saez, Pepe Aróstegui, Helen Simons, Rob Walker, Clement Adelman, Maria Bustelo, Teresa Vasconcelos, Fátima Cruz, Gloria Contreras, Ivan Jorrin, Jimena Tirado, John and Christine Elliott, Mike and Ann Atkin, Dick and Jenny Harvey, Gordon Hoke, Del Harnisch, Edith Cisneros, April Munson, Yujin and Nettie Lee, Luisa Rosu and Ana, Catalina and Louis Ulrich, Lizanne DeStefano, Kathy Ryan, Joyce Grant, Claryce Evans, Ben Stake, Sara Stake, Ben Joselyn, and of course, Bernadine.

Criticism and the University

We have been graciously convened today by Rector Evarísto Abril Domingo, here at the University of Valladolid. This a historic place. By some accounts, the University of Valladolid is the eleventh oldest university in the Western Hemisphere, a teaching center before 1200 AD.

ALIENOR REGINA
ALFONSUS REX
MAGISTER

Just 800 years ago, the King of Castile and Leon was Alfonso the Eighth. He is in the middle of the picture here with Queen Alieanor on his right. Alfonso created the University of Palencia from a cathedral school and studios at Palencia and Valladolid.[1] I count him my grandfather, 22 generations removed. I trust he is here in spirit with us today. Perhaps you are skeptical. A theme of mine is that it is good to be skeptical.

According to Historian Olaf Pedersen,[2] the function of the first universities was quiet scholarship and teaching, a small number of professors teaching doctrinal-knowledge for church and kingdom and law and the market forces of those times.

[1] Pedersen, O., 1997. *The first universities.* Cambridge University Press.
[2] Ibid.

With the teaching came advocacy, mostly advocacy of the traditions, the *status quo*, and advocacy for the quelling of heresy. Both advocacy and heresy enlist criticism. Criticism was present at the forum in Athens, at the tutorial in Bologna, and at each university from the beginning.

We should be skeptical. Skepticism is a disposition to doubt the story being told, or to doubt explanations in general. Philosophers have examined doubt and skepticism and, in the extremes of Cartesian doubt, found it leading to deep discouragement. But – offered as the "suspension of belief while testing that belief" -- doubt energizes scholarly reflection. In "The Agony of Christianity," Unamuno[3] said, "Faith without doubt is dead faith."

Criticism

Criticism is the expression of doubt, opposition to explanation and performance. Criticism is a human disposition, not an invention of modernity. It is personal more than institutional. It is possible that skepticism was prominent in the teachings of the universities of the Middle Ages. But pride in criticism, pride in critical power, was not their reputation[4] Science emerged, and sometimes expressed itself in criticism. But it was not until the Nineteenth Century that even research, let alone criticism, became formalized as a responsibility of university study[5].

In Academia, criticism is usually thought of as addressed to a document or doctrine. More often than not, criticism is an expression of opposition and fault-finding. In the construction

[3] Unamuno y Jugo, M., 1974. *The agony of Christianity.* Princeton University Press.

[4] Pedersen, O., 1997. *The first universities.* Cambridge University Press.

[5] Frängsmyr, T., 2006. Universities, research and politics: The avoidance of anachronism. In K. Blückert, G., Neave, G., & Nybom, T., editors, *The European Research University*. New York: Palgrave Macmillan.

of a new theory, there is an inherent holding other theories as inferior. Criticism is often used to protect the *status quo*. In teaching and inquiry, new ideas emerge that contrast with the old, and criticism occurs whether or not the researcher intended to be contentious.

The complexity of criticism is illustrated in the life of Benjamin Jowett of the faculty of Oxford University in the Nineteenth Century. Though a man of the cloth, Jowett was publicly seen as a critic of religion. But no! In 1841, at Westminster Abbey, Jowett's sermon claimed that liberal interpretation of the Bible promoted new understandings of nature and greater appreciation of the complex world God had created. Jowett said,[6] "The criticisms of the present day will at first be felt as a blow to faith, but they will issue in fuller establishment; all that is important will survive."

My concern today is how criticism is treated in the policies and practices of the university. I will speak of a university's obligation to support challenges to accepted truths. I will use my own field, that of educational program evaluation, to illustrate the need and obstacle to challenging conventional arts and ordinary science.

Evaluation

Evaluation is the recognition of quality. Evaluation is seen as the discovery of goodness, and as the discovery of human interpretation of goodness. Evaluation is present in all human behavior, but is formalized and professionalized with the specification of an evaluand (the thing being evaluated), and

[6] Jowett, B., 1871. Darwinism, and faith in God. Published in Sermons on faith and doctrine. London: John Murray.

specification of criteria, standards, and critical incidents, by which merit is recognized.

Through the Dark Ages, the universities served the church and feudal powers to indoctrinate newcomers in the knowledge needed to carry on convention and proclamation. University professors advocated doing things right and they often had the discriminative power to recognize what was flawed and heretical. The professors were evaluators. They had critical power. Across several centuries, at least from our point of view, that evaluative power was seldom exercised in reform or problem solving or extending social well-being but used more to maintain the security of church and kingdom.

Following the Middle Ages was a time of new interpretation, new expression; and after the Reformation came a counter-reformation, dominated, particularly in Spain, by the Inquisition, a stunningly brutal exercise in criticism.[7]

The Spanish Inquisition was a royal tribunal established by King Fernando the Second of Aragon and Queen Isabella the First of Castile, intended to maintain Catholic orthodoxy, particularly among converted Jews. The Inquisition, in Spain and elsewhere, was one of the most serious exercises of evaluation in Western history. It drew its criteria and standards from interpretations of sacred writing.

The Inquisition was gradually pushed aside by a new heresy, Science. Still it lasted into

[7] Netanyahu, B., 1995. *The origins of the Inquisition in Fifteenth Century Spain,* Second edition. New York Review of Books.

the Nineteenth Century. Science also was fundamentally an act of evaluation and criticism, drawing on evidence largely from observations and causal reasoning.

Human Perspective

We find some contemporary authors saying that educational program evaluation began after World War II, but, like most early universities, it had no discrete birth or invention. Informal evaluation is a fundamental part of living, of experiencing, of worship and science, of civil and lay society. In the last half of the Twentieth Century, formal evaluation did change considerably, as a sign and as a determining force of modernity. We codified, routinized, technologized and explicated, even more than during the Inquisition. We put more labels on things and people. We raised claims and expectations that evidence would be gathered to justify collective action. Fifty years ago, a few people invented the name *program evaluation* to facilitate the contracting of a formal search for quality in social and educational programs. Most people continued assessing their programs without realizing the name had changed.

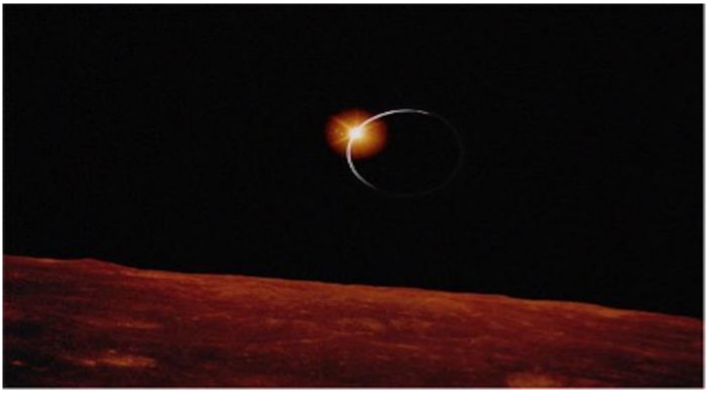
It is not a new discovery that people see things differently. Drawing his text from the Bible, Benjamin Jowett[8] opened a sermon urging respect for Charles Darwin with: "*The sight of nature affects men differently in different ages and countries. We ourselves receive different impressions from natural scenes when the sun shines upon them and when they are enveloped in mist and storm; and our perceptions of them also vary with the varying moods of our own minds.*"

[8] Jowett, B., 1871. Darwinism, and faith in God. Published in *Sermons on faith and doctrine*. London: John Murray.

Guernica. Picasso, 1937.

A fire in the fireplace looks not the same in winter and summer, nor does a seaside beach. Even when the physical is held constant, different viewers see Picasso's *Guernica* differently. Is the right way to see *Guernica* how Picasso saw it?

Some critics, some evaluators, say, "Yes, the Creator should be served." But in qualitative evaluation studies, there is advocacy that even the most unusual interpretation will enrich the understanding of the evaluand. There are many right ways to see a sunset.



Sunset Photographed from the Moon

Honorary Doctorates

According to the *Oxford Dictionary of National Biography*, the first known honorary degree was bestowed in 1470 on Lionel Woodville by Oxford University. He had been elected Chancellor, holding only a baccalaureate degree, and perhaps was deemed worthy of higher qualification. In 1483, as a participant in a rebellion against King Richard the Third, Woodville was deeply engaged in criticism. The University chose not to stand with him and began looking for a new Chancellor.

The first honorary degree by the University of Valladolid was awarded in 1964 to Sir Charles Alexander Petrie. According to *Wikipedia*, Petrie, an Irishman living in England, was a popular writer on monarchism, giving some attention to my possible grandfather, Alfonso the Thirteenth and others of the Spanish Royal House. Petrie was not uncomfortable with Fascism. To some, he was an apologist and appeaser. Clearly he was engaged in criticism. He wrote about the failure to restore the Stuart line, the Catholic line, to the British throne. Would it not have been interesting to sit in on the meetings where Charles Petrie was evaluated for selection for *honoris causa*!

There have been 72 honorary academic doctorates granted on this campus. One was to the distinguished historian of language, Antonio García-Berrio of the University of Complutense. García-Berrio was a scholar of criticism. He claimed that the language of literary criticism was insufficient for art criticism.[9]

And Barry MacDonald. Several of us here today were here also in 1999 when the Scotsman MacDonald was awarded an honorary doctorate. As I see it, he was honored for his criticism. In our field of program evaluation, he had experienced the power of central governments, corporations and foundations to constrain evaluation to what served bureaucratic and economic purpose. He led the way to a more democratic approach to the evaluation of curricula and training. His life has been spent in

[9] García-Berrio, A., 1992*. A theory of the literary text.* Berlin: Walter de Gruyter.

inquiry and professional practice, largely responding to the appetite for control by governments. His university did not celebrate MacDonald's development of critical power, but Valladolid did. Across Spain, MacDonald's challenging ideas have been honored, and he was, as I am today, invested with the honor of a lifetime, the Valladolid honorary doctorate.

Evaluation throughout Scholarship

I have mentioned three earlier recipients of the *honoris causa* of this university. I bring to your attention the critical power of their work. Now I am going to claim that whether the scholar distinguishes himself or herself as a creator or as an analyst of ideas, he or she has worked deeply into the discipline of evaluation. Certainly not all evaluators are scholars, but all scholars are evaluators.

The many fields of the university are broadly represented among the honorary doctorates since 1964. Beyond the three I mentioned, here are six more: Severo Ochoa, biochemistry; Valentín Fuster, cardiology; Benzion Netanyahu, history; Michéle Aymard, zoology; Mario Benedetti, literature; and Antonio Fernández, architecture.

The disciplines of the campus have become so specialized that these six scholars could have sat at the same table at the Residencia de Estudiantes and talked congenially -- and gained little vision of the intellectual territory each other has advanced. But their skills of navigation across that territory have at least one more thing in common, they all have evaluated: *where* the mind has been before, the options before them, the suitability of their methods, the disappointment of their mistakes, ... evaluated.

Evaluation is the search for quality, not only the quality of sketches, theories, and surgical procedures, but the quality of each line drawn, each relationship calculated, each stroke of the knife. Each step forward is an *evaluand*, held to criteria, searched for implication, reviewed with critical eyes, felt with haptic senses. The scholars will not necessarily boast of their skills as evaluators, but with every breath they are alert to the appearance of fissure and incompatibility.

Many of us in program evaluation squint to catch the vision of Michael Scriven, a philosopher from Oxford turned evaluation theorist. It is he who has spoken most eloquently about the ubiquity of evaluation across the humanities, sciences, and professional studies. Scriven said:[10]

> *Just as scholarly disciplines share in the use of language, ethics, and aesthetics, they share in using (and misusing) the processes of evaluation. But unlike those other processes, across the colleges, there is no central focus, no discipline, for the study of evaluation. The time has come to acknowledge such a discipline.*

Evaluation as a Discipline

In the early 1960s, the field of formal educational evaluation was not yet a field. It was an offshoot of measurement in education. There were practical uses of measurement for such questions as trainee readiness for training and expression of the goals of teaching, but the credentials of evaluators then were the credentials of educational assessors. And not surprisingly, evaluation was conceptualized then -- and often still -- as a part of the social

[10] Scriven, M., 1996. Evaluation, the skeleton in the disciplinary closet. Millercomm Lecture, April 23. University of Illinois.

science of teaching and learning. The leading measurements theorist of the day, Lee Cronbach,[11] urged that program evaluation be designed to assist educational research and development. And many agreed.

Michael Scriven said,[12] "!Suficiente!" He said, "Enough of that!" He urged that evaluation be aimed at the consumers, mainly the professionals, the public, the policy makers, those who needed to know the quality of teaching and learning.

In the 1960s, standardized testing in America was used largely to assist academic counselors to help students and parents to make good scholastic choices, particularly with regard to choosing courses of study and colleges. But the politicians and economists of America, and soon the world, saw that testing could be used to control education and, diabolically, to obstruct the investments in it.

Test-based assessment became the evaluator of schooling. Some professionals bought into it, many did not, but evaluation has been greatly constrained by the unwarranted belief that the quality of schooling is well indicated in the scores of students on standardized achievement tests. The minds of the educational world today are captured by a political-economic illusion, an illusion that simple criteria (such as PISA and TIMSS scores) can represent the quality of education.

The Silent University

Over the centuries, the Western universities have operated with considerable autonomy and in considerable silence. By and large, they have operated in harmony with royal,

[11] Cronbach, L. J., 1963. Course improvement through evaluation. *Teachers College Record, 64*.

[12] Scriven, M., 1996. Evaluation, the skeleton in the disciplinary closet. Millercomm Lecture, April 23. University of Illinois.

state, and public tastes, and with chancellors selected with those tastes in mind.[13]

A few faculty members will speak out critically of campus practice and disparagingly of social neglect. A larger number of students will protest. But seldom can they lay out a "university position." The tradition of academic freedom is maintained, not an urgency of doing better, and the voices are weak. The universities have a platform to speak critically, but mostly remain silent.

Who could speak for the university? There are many voices and none. The Ministry of Education, the Rector, the governing boards, the faculty senate, the student leaders? No, almost none can speak for the university. They speak as individuals. They identify with the university -- but seldom lay out the university's moral position. The universities barely murmur.

But even if silent, each of these many university people contribute to an ethic of, to some extent, promoting and, to some extent, constraining criticism. Few speak out when those around them, those respected and held dear, caution against speaking out. The leadership of the university, one and the many, has potential for creating an atmosphere of social, political and academic criticism. Largely an unused potential.

The Bologna Process

When the university itself is being changed, there is great need for criticism from within and 'round about. Such is the case with the Bologna Process. The Bologna Process is a complex development for improving European higher

[13] de Ridder-Symoens, H., 2003. A history of the university in Europe, 2. *Universities in Early Modern Europe, 1500-1800.* Cambridge University Press.

education,[14] Common standards were initiated in 1988 with the Magna Charta Universitatum and in 1990 with the Tempus Programme (the Trans-European mobility scheme for university studies) and in 1998, the Sorbonne Declaration. Then in 1999, the Ministers of Education of 29 countries met in Italy to begin changes in the degree-granting functions of participating universities so that their courses would have greater common meaning and allow more appropriate exchange of student credit. Support came from the European Union. And expectations grew. Sheffield University Professor Ann Corbett said:[15]

> *The aims [of the Bologna Process] are external to Europe, and internal. The goal is not only to make the European higher education area (EHEA) attractive enough to the rest of the world to draw in more of the best foreign students and scholars, but also to boost quality within Europe itself, as a way of making universities more effective within the knowledge-based economy which the world's richest nations regard as the* sine qua non *of economic growth.*

The Bologna Accord included reasoned purposes, but posed difficult tasks. And the tasks grew as new aspirations arose. According to Education writers, Madeleine Green and Andris Barblan:[16]

[14] Zgaga, P., 2006. Looking out: The Bologna Process in a Global Setting. Oslo: Norwegian Ministry of Education and Research.

[15] Corbett, A., 2005. *Universities and the Europe of Knowledge, Ideas, Institutions and Policy Entrepreneurship in European Union Higher Education Policy, 1955-2005.* Houndmills: Palgrave Macmillan. p xii, 4.

[16] Green, M., & Barblan, A., 2004. Higher education in a pluralistic world: A transatlantic view. Washington: American Council on Education. page 6. www.cua.be/cua/jsp/en/upload/Translantic_Dialogue_2003.1 12920893.pdf

In spite of claims that the process was increasing the cultural and linguistic diversity of European higher education institutions, the most visible energies have been to increase standardization of teaching and curricula and assessment.

For the idea of interchangeability of universities, there has been strong support among political leaders and economists.[17] Given the size of changes undertaken, the criticism of the Bologna Process has been small. Chris Lorenz of the Free University of Amsterdam did speak out:[18]

The basic idea behind all educational EU plans is economic: the basic idea is the enlargement of the scale of the European systems of higher education, ... in order to enhance its 'competitiveness' by cutting down costs. Therefore "a Europe-wide standardization of the 'values' produced in each of the national higher educational systems is called for." [The Bologna Process] proposes educational reforms that [would] erode all effective forms of democratic political control over higher education. "It is obvious that the economic view on higher education recently developed and formulated by the [European Union] Declarations is similar to and compatible with the view developed by [the World Trade Organization: WTO] and [the General Agreement on Trade in Services: GATS] the Dutch universities and the Bologna Process. [19]

The processes have been discussed and debated mostly in the privacy of administrative circles. But the faculties,

[17] Zgaga, P., 2006. Looking out: The Bologna Process in a Global Setting. Oslo: Norwegian Ministry of Education and Research.

[18] Lorenz, C., Will the universities survive the European integration? http://dare.ubvu.vu.nl/bitstream/1871/11005/1/Sociologia%20Internationalis.pdf

[19] *http://dare.ubvu.vu.nl/bitstream/1871/11005/1 /Sociologia%20Internationalis.pdf*

students, and public have let it happen quietly. They have not been abundantly informed nor invited in for criticism.

Protecting Human Subjects

Let me speak briefly of another internal university matter that has lacked adequate criticism. In America and around the world, for more than 20 years, efforts have been made to reduce the risk to human subjects from university research. On my campus, as on all in the U.S.A., we require formal research plans to be reviewed to reduce likelihood of physical and social injury to individuals participating in the studies.

The purpose is humane, the advocacy rational, and participation was insured by threat that federal funding would be held back for universities not complying. Unfortunately, a narrow standard for good research was specified and faith was placed in researchers to follow their proposals closely. Protests against "mission drift" (over-reaching their commission) and presumption that all disciplines are the same, these protests have been voiced, as here by sociologists Charles Bosk and Raymond De Vries:[20]

> *The problem of [Institutional Review Boards] for qualitative research is that they are such a distraction from the real difficulties that we face and from the real ethical dilemmas that confront us, that we may not recognize and discuss the serious and elemental because we are so busy with the procedural and bureaucratic. (p 99)*

Criticism can be found, but the 500-some American universities have seldom encouraged critical reviews of human

[20] Bosk, C. & De Vries, R., 2008. Bureaucracies of mass deception: Institutional Review Boards and the ethics of ethnographic research. In *What would you do?* University of Chicago Press.

subjects review. The result is that researchers young and old are discouraged from using situation-sensitive methods and avoid research questions that sometimes require months of review by the reviewers.

Another matter of importance, unchallenged by the universities, is the evaluation of Third World development projects such as with World Bank and the International Monetary Fund. Many of the studies are evaluated in a perfunctory way by people whose income depends on remaining in good standing with those agencies, seldom submitted to meta-evaluation or public review. American universities are the main custodian of research standards in America, but they are not vigorously reviewing this questionable evaluation.

Character of a University

Criticism sometimes requires bravery. Extraordinary was the bravery of Galileo, Joan of Arc, Darwin, García Lorca, Severo Ochoa ...

According to the "diamond theory" of ancestral lineage,[21] in the year 1150, I had more than 20 million ancestors. In other words, I had less than one chance in 20 million of inheriting bravery from Alfonso the Eighth. How can I be brave unless my university has the character to draw out my bravery?

What defines the character of a university? It is not as much in the awards it receives as in the awards it gives. Character is defined by the scholarship of its teachers and students. Character is defined in the leadership of its administrators. Character is defined in the expectations of citizens, legislators, teachers and students. All those people are

[21] Manrubia,, S. C., Derrida, B.,& Zanette, D. Z., 2003. Genealogy in the era of genomics: *American Scientist,* 91, p. 158-165.

evaluators. They have standards against which they hold their university.[22]

Criticism remains a poorly thought-out standard for universities. Social criticism is low priority. The trustees, the accrediting agencies, the faculties and students make little demand that universities promote the humanities of criticism. Criticism is expected within each discipline -- but criticism of the university itself is not welcomed. It loses friends. Criticism offends those who have invested their lives, their savings, their children in the university. It hurts many to hear the university criticized.

But today's university is part of a society impoverished of serious self-evaluation. The sign in the window should say, "Aqui se hace crítica." And the university needs be more than a site where criticism is heard, more than a cradle of criticism. It should increasingly be a host to criticism, a nourisher of criticism. The university should be a generator of critical power.

The university should be steadfast in warning that "market forces" can silence criticism. Stig Strömholm, *rector magnificus* of Uppsala University, said[23] that the obligations of a university included making: "... all students aware of the fact that it is an institution with a responsibility [for] frequently unpopular and controversial self-reflection [on] the community to which it belongs."

To conclude:

[22] Nybom, T., 2006. Creative intellectual destruction or destructive political creativity? Critical reflections on the future of European "Knowledge Production." In Blückert, K., Neave, G., & Nybom, T. editors, *The European Research University*. New York: Palgrave Macmillan, p 3-16.

[23] Strömholm, S., 2006. Summing up. In Blückert, K., Neave, G., & Nybom, T. editors, *The European Research University*. New York: Palgrave Macmillan, p 177-181.

When the University of Valladolid and all the early universities began, they had the power to limit the spread of knowledge. Today, it is their business to spread knowledge, experience, and inter-reliance -- to the far corners of the earth. To some extent, information has been democratized.

But corrupt and insidious information flows abundantly and enticingly. Information needs evaluation. Advocacy needs scrutiny. Universities should take as much pride in their critical power as their Nobel Prizes.

In the years ahead, there will be much need for critical power. More help from the universities is needed. Thank you for hearing my words. Thank you, University of Valladolid.

2011

On campus, I went to a lecture by Ruth Watkins, President of the University of Utah. She spoke ardently about business and industry having a role in education research and development on campus. I worried once again about the research questions being pursued and the criteria of merit for the studies.

Graduation Rates

The choices of research questions made by researchers themselves have always been imperfect. They study what they study because it is what can get funded, because it is likely to be published in the better journals, because it looks good on resumes, because it is fun, because it fits their biases. And seldom because it has any real chance of extending our understanding of disciplinary knowledge or provides understanding for improving human services.

We are fortunate that we have few administrators or institutional research offices pressuring researchers on what to study. A few people or corporations who make large gifts do express appreciation for certain research and development. Some administrators and some donors sometimes exercise some influence on how faculty work enhances the institutional reputation. And reputation and human service are not perfectly correlated. It is good that researchers have as much freedom as they have.

Of course, corporate values are not entirely perverse, trivial, or self-serving. Collaboration will sometimes make the studies better. Simplicity may be needed. Collaborators won't necessarily be manipulative. But most of them, I think, are serious about their business plans, their public standing, even their own advertising. And researchers do not want their next study to hurt prospects for another after that.

We can expect the criteria for evaluating research to continue to move away from what provides needed knowledge and support for human living. Take one instance mentioned by President Watkins: graduation rate. She indicated that the University of Utah was not doing as well as the University of Illinois to assure that every student graduates.

Graduation rate has been pushed by business-oriented McKinsey and Company as an indication of quality. Many central administrators in the Chicago community colleges, for example, but also on the Illinois campus, have pushed for emphasis on degree programs. It will give more thrust to what some students' study. And it may facilitate control of department offerings. But such degree-completion orientation discriminates against part-time students. It further commodifies what students study. And it will not encourage admission offices to admit students on the basis of who can best profit from the experience.

Who are most likely to be admitted? The youngsters already advantaged and unlikely to drop out. Who are most likely to be turned down? The poor. The self-motivated. Those with disabilities.

In choosing simple criteria of merit and popularly understandable designs of research, the research of the university might become better funded, and even more democratic, but may become less likely to do what it could for advancing disciplinary knowledge and grounds for improving human services.

2013

It has not been uncommon for advanced students, as well as those just beginning, to ask for clarification as to how to do research.

Freedom from the Rubric

During World War II, President Franklin Roosevelt called for commitment to four freedoms: Freedom of speech, freedom of worship, freedom from want, and freedom from fear. The tones of his expression still bring tears to a few older eyes.

They were war-ravaged years, now passed almost beyond memory into a more complex world. We have other problems. One, little recognized, calls out for another freedom: Freedom from the rubric.

In order to evaluate complex matters in our lives, many of us borrow or compose a rubric. A rubric is a set of criteria, a list of essential characteristics, for grading things. It is a reminder of what we have thoughtfully considered the essence of quality.

A rubric is a set of red-letter ideas, important enough to be the titles of our chapters. President Roosevelt had a rubric for a world free of war and oppression. *His* rubric had four essentials.

A rubric itself is not necessarily good. It may be overly simple. A rubric is not necessarily bad. Sometimes greater conformity to explicit standards *is* needed. The rubric itself and its use needs to be evaluated.

One rubric for "good expression," a rubric for teachers of writing, draws our attention to: content, organization, word choice, voice, fluency, and convention. According to the rubric, organization includes "good transition" from one part to the next.

The literary world changes too. What was once a standard for expression, may *not* be essential for what we do now. Beethoven sometimes avoided transition, as does Steven Spielberg.

A rubric helps one think about some things and keeps us from thinking about others. A rubric does not read the essay, and then think, "What is good here?" The rubric usually looks for deficits. Are these six things weak or missing here?

The human mind is capable of looking at a writing sample or at a teacher teaching writing -- and finding strengths and weaknesses, or more. The mind is capable of thinking of holistic quality as separate from the goodness of parts.

A rubric standardizes the evaluating, but does not necessarily make the evaluating better. As Elliot Eisner said[1], evaluators need to be connoisseurs, using all their sensitivities and experience to examine the evaluand at hand.

In some places, the rubric has attained something of a dictatorial power. It sometimes usurps the laws of experience. This may be time to battle for freedom from rubrics.

[1] Eisner, E. W., 1979. *The educational imagination: On the design and evaluation of school programs.* Macmillan.

2018

When are there not more degrees of right and wrong? Is arbitrary best learned by experience?

Right and Wrong

We teach what is right and we teach what is wrong. And we teach some of what used to be wrong is no longer wrong. And right right. Yes, we do tell children to sing and spell correctly and not to lie, but we don't teach much about what is wrong about wrong. The right spelling of color is c-o-l-o-r, except a wrong spelling is right in some places. We do lie by telling sick people they look good when they don't. Doesn't Ella sing flat? Some time it's right to lie. Wrong can be kinda wrong or really wrong.

I have been working on my family tree. I have been told that in Amberson Valley, Pennsylvania, the birthplace of my father's father, Albert Stake, that there were three families of Stakes there in that small valley, the three unrelated. But surely they were related somehow sometime. It is too much a coincidence that three families with the same unusual name settled in the same valley. But of course, one family could have been the offspring of an adopted child. And one family could have been the offspring of a long-time hired hand who for some reason acquired the name of his hospitable master. And other possibilities lie close.

As I look at the hundreds of marriages on the chart, I wonder how many of them really show the parents of that child in the next generation. If we were to test all DNAs, would we find 2% of the links to be wrong, or 8%? We have some doubts about census-taker spellings and numerals in family Bibles. And someday will we question DNA? Genealogists have standards of evidence, but truth remains uncertain.

Historians have interpretations and scientists have hypotheses. What is taken as truth is agreement among those who command the evidence, which is likely to stand 'til evidence or scholars reconstitute. There is quite a bit of arbitrariness as to what is right and wrong.

My visiting saxophonist Clement Adelman lay awake listening to the wind chimes in our tree, compelled to compose a second line for each successive tolling. Were there no wrongs? Are all random lines right for rhapsody? How can I hear the Beatles so right now when they were so wrong for me in 1960? Alex, is rap music?

Perhaps the reason we are reluctant to teach about the nature of evidence is that almost always it will be circumstantial. Is it not true that there is always a somewhere for which that evidence is sufficient? For every hanging judge is there a forgiving jury?

Music teachers have you kids play trumpets by the score. And correct every misplay. Ever, ever reverting to the score. Or shall we sometimes stop and ask under what circumstance this might be the succession. Does this not stretch the envelope? At what length will the canon evolve?

We have a curriculum built upon what is right, given that some (with little caprice) have decided it to be not wrong. Our curriculum rings with authority. We don't much want students to expect to figure out what is right. Is it because sooner or later that will leave us wrong? We treasure compliance. We have a large investment in our standards.

Or is teaching the right, the canon, the efficient way, toward a better right, the next canon? We can count on error. We can count on learners to judge. We can count on an allure to the alternative. But surely jazz was not worth the evil of slavery.

Neither truth nor error is pure. The mixture will vary from place to place. The mix will be encountered from nursery to classroom to concert hall. We postpone the effect with lock steps and common goals. And are caught in a consummate effort to preserve the curriculum we inherit.

2018

*Instead of story-teller, I could call him a story-*évocateur*. He was an excellent interviewer. He got people to answer a few questions, then pretending he didn't understand all they meant, (What did you mean by that?) got them to slide from explanation to ethnography.*

Terry Denny, Story Teller

Terry Denny was a good story-teller. He found good stories to tell. Sometimes, fish stories. You can't be a good story-teller without good stories to tell.

Where did Terry get his stories? Not at the bar. Not from imagination. He got good stories by being a good observer and interviewer. I watched him teach students to do interviews. His best questions did not seek information. They sought glimpses at the turmoil in the interviewee's mind. He looked for what the interviewee might least likely tell him.

One of his favorite questions was, "What did you really mean by that?" He wasn't so much looking for the meaning of the previous utterance, but for what the interviewees would do when thinking they had been challenged. He knew it is stressful when one needs quickly to find something plausible to say.

Interviewees seldom resisted. Somehow they felt that Terry was on their side. They liked Terry. They could trust Terry. Many times they wanted to say something that Terry could use. Maybe the truth. Maybe something tied to a story.

I remember Terry's license plate. Not the last one, the earlier one. It spelled: X P 8. Expiate. Expiate. What does expiate mean? If you are Catholic, you know. They tell me it means "Atone for your sin." "Pay up." Confess. Tell me your story. Expiate.

Mean questions? I've asked my share of mean questions. I sometimes realize later, ooow, that was intimidating. Terry didn't ask intimidating questions. Some were tough but more likely asking for experience, clarification. "What did you really mean by that?"

They loved Terry, and told him secrets. Terry sometimes worked alongside Klaus Witz. He admired Klaus. Klaus used a kind of case study called "portraiture." By repeated, intensive questioning, Klaus could get pretty close to something like the essence of personality, the deeper values. Terry saw this kind of questioning not so much getting a personality profile but more an opening the door to personal stories, the most important stories to tell, or not to tell. Terry could get them to tell.

Terry would sometimes liken research to fishing. How are you going to catch the big one? He saw fish as having different moods, different places to hide. He was less interested in a *Hummingbird Fishfinder* than a topographic map of the bottom, telling him where that fish might hide. What lure to use needed to match the mood of the fish. So finding the story was looking for places that memories might hide.

No, no. I'm wrong in implying that guilt is the rock behind which the story hides. Here's a bit from Terry's book, *Being with the Dying*[1]. It's about Art, a taciturn fellow, an older man.

When I finally got past being preoccupied visually with Art's behemoth physical frame, I discovered a handsome face, graced by long hair, peppered gray, and lively gray eyes. His beard and mustache were trimmed, nails clean -- and his personal hygiene was fine.

[1] Denny, T., 2015. Being with the Dying. Mahomet, Illinois: Mayhaven Publishing.

Our teeter-totter chats began with a large man using guarded, short sentences on one end, with a skinny man talking too much on the other. As the weeks went by, Art's sentences lengthened and mine shortened. He had a surprising willingness to share some of his carefully crafted views of the world.

Terry wrote that Art was a voracious reader, that he was fascinated with space and time. Their conversations eventually ranged from how to tie an Adams trout fly on a #12 hook, to an ignominious aspect of personal space whenever an arrant BM missed the mark in a small trailer. "But all that and more was yet to come."

It wasn't so much that the story was there for the telling. The story was there for the trolling**.**

A kind of one.

2018

The first version of this paper was written to be a sermon at the Unitarian Universalist Church in Urbana. Then I gave a variation each to Saville Kushner and Merel Visse for their blogs. The idea came to me in a dawn reverie and seemed not likely to be right because I couldn't remember reading it elsewhere. Alternate title: Talent, Equity and Ranking.

Those Not Chosen

Even the poorest of us has the power to allocate privilege. With warm smiles and sincere attention, we help make other lives livable. With carelessness and honor roles, we make other lives less livable. Our social policies work to make life more livable for some children, and less livable for others.

I want to speak about how scholastic views of talent, equity, and course-grading influence the self-perceptions of children and all of us. I base my claims on experience as a teacher and psychometrician and evaluator of teaching. In my sixty years of practice, I did not gather hard data to support these claims. But I believe what I say. Standardized testing and teachers testing needlessly facilitate an academic caste system.

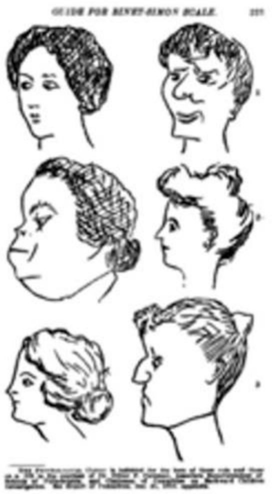
Talent

As I mentioned in 1984, in 1930 my Mother took me to the Nebraska State Fair and entered me in a baby contest. I won a prize. The announced purpose of the contest was to examine the health of drought-years children, mostly rural, many not being seen regularly by a doctor.

At the nearby Kansas State Fair, the other purpose of the examinations was framed on the clinic wall, something like: "...

to discourage breeding in families of inferior talent." The international Eugenics Society, I believe, provided the posters. That was about the time Adolph Hitler was running for Chancellor of Germany.

The societal measurement of talent, particularly intelligence, even more particularly, scholastic aptitude, began in Paris just over 100 years ago. Alfred Binet, a psychologist studying mental function, was asked by the French government to create a test to discriminate between children able to learn in school and those unable.



One of the items of the Simon-Binet test is shown here, asking, for each pair, "Which of these two faces is the prettier?"

With my awfully-pretty Bernadine and four talented children, in 1963, I moved to the University of Illinois. I was the new Assistant Director of the Illinois Statewide Testing Program. We provided standardized tests to some 450 Illinois school districts, not including Chicago.

I was happy in the job, believing that those standardized tests were a search for talent, a finding of youngsters who weren't doing well in school, but would develop skills, intellectual skills, with less standardized curricula.

At the time, the tests were designed to help school counselors give guidance to youngsters. Not yet were they intended to improve curricula, nor to evaluate teaching, nor to compare schools. Not yet were the tests paid attention to by politicians.

In 1960, a testing man from the University of Pittsburgh, John Flanagan, created Project Talent.[1] Testing 400,000 kids nationally, he was seeking hidden talent among the nation's youth. But -- as so many social science studies go -- the interest in identifying actual children gave way to interest in finding correlation of variables. The research mostly served additional research. Perhaps this work should have been in the Humanities but was located in the Social Sciences.

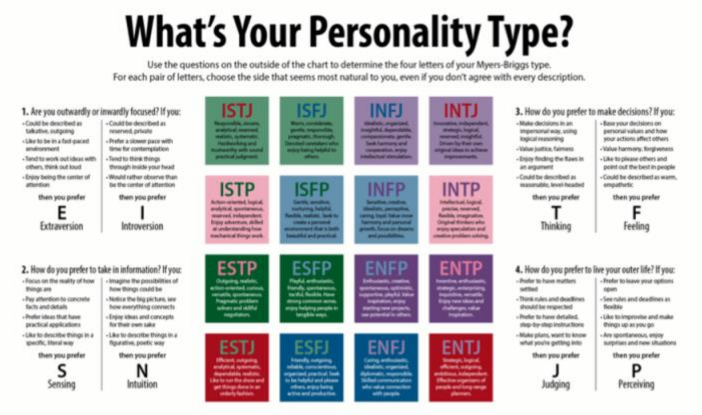
Equity

I had had graduate work at the Educational Testing Service. The President of ETS was Henry Chauncey. Although ETS became wealthy and famous, with the Scholastic Aptitude Test, by measuring verbal and mathematics learning ability, Chauncey was embarrassed by the narrowness of this definition of talent, and urged his researchers to search far and wide.

At its beginning, around World War Two, ETS marketed course-specific achievement tests: geography, algebra, penmanship, biology, … but, one test fits all, they didn't fit the diversity of schooling in the nation's classrooms.

[1] Flanagan, J. C., Davis, F. B., Dailey, J. T., Shaycoft, M. F., Orr, D. B., Goldberg, I., Neyman, C. A., Jr., 1964. Project Talent. University of Pittsburgh, Project Talent Office.

At one research meeting, I listened to a discussion of a new personality inventory, the *Myers-Briggs Type Indicator*, authored by two women, Katherine Briggs and Isabel Myers. It was based on the Jungian idea that persons were highest on one of four psychological functions: sensation, intuition, feeling, or thinking. Measuring *these* functions might lead to better recognizing the complexity of talent and aspiration and the opportunity for more individualized teaching. But key researchers at the meeting were adamant in their assessment that the validity of *Myers-Briggs* was low. Their question was not, "Would it help?" but "Is it precise enough?"



What's Your Personality Type?

Use the questions on the outside of the chart to determine the four letters of your Myers-Briggs type.
For each pair of letters, choose the side that seems most natural to you, even if you don't agree with every description.

1. Are you outwardly or inwardly focused? If you:

- Could be described as talkative, outgoing
- Like to be in a fast-paced environment
- Tend to work out ideas with others, think out loud
- Enjoy being the center of attention

then you prefer
E
Extraversion

- Could be described as reserved, private
- Prefer a slower pace with time for contemplation
- Tend to think things through inside your head
- Would rather observe than be the center of attention

then you prefer
I
Introversion

2. How do you prefer to take in information? If you:

- Focus on the reality of how things are
- Pay attention to concrete facts and details
- Prefer ideas that have practical applications
- Like to describe things in a specific, literal way

then you prefer
S
Sensing

- Imagine the possibilities of how things could be
- Notice the big picture, see how everything connects
- Enjoy ideas and concepts for their own sake
- Like to describe things in a figurative, poetic way

then you prefer
N
Intuition

ISTJ Responsible, sincere, analytical, reserved, realistic, systematic. Hardworking and trustworthy with sound practical judgment.	**ISFJ** Warm, considerate, gentle, responsible, pragmatic, thorough. Devoted caretakers who enjoy being helpful to others.	**INFJ** Idealistic, organized, insightful, dependable, compassionate, gentle. Seek harmony and cooperation, enjoy intellectual stimulation.	**INTJ** Innovative, independent, strategic, logical, reserved, insightful. Driven by their own original ideas to achieve improvements.
ISTP Action-oriented, logical, analytical, spontaneous, reserved, independent. Enjoy adventure, skilled at understanding how mechanical things work.	**ISFP** Gentle, sensitive, nurturing, helpful, flexible, realistic. Seek to create a personal environment that is both beautiful and practical.	**INFP** Sensitive, creative, idealistic, perceptive, caring, loyal. Value inner harmony and personal growth, focus on dreams and possibilities.	**INTP** Intellectual, logical, precise, reserved, flexible, imaginative. Original thinkers who enjoy speculation and creative problem solving.
ESTP Outgoing, realistic, action-oriented, curious, versatile, spontaneous. Pragmatic problem solvers and skillful negotiators.	**ESFP** Playful, enthusiastic, friendly, spontaneous, tactful, flexible. Have strong common sense, enjoy helping people in tangible ways.	**ENFP** Enthusiastic, creative, spontaneous, optimistic, supportive, playful. Value inspiration, enjoy starting new projects, see potential in others.	**ENTP** Inventive, enthusiastic, strategic, enterprising, inquisitive, versatile. Enjoy new ideas and challenges, value inspiration.
ESTJ Efficient, outgoing, analytical, systematic, dependable, realistic. Like to run the show and get things done in an orderly fashion.	**ESFJ** Friendly, outgoing, reliable, conscientious, organized, practical. Seek to be helpful and please others, enjoy being active and productive.	**ENFJ** Caring, enthusiastic, idealistic, organized, diplomatic, responsible. Skilled communicators who value connection with people.	**ENTJ** Strategic, logical, efficient, outgoing, ambitious, independent. Effective organizers of people and long-range planners.

3. How do you prefer to make decisions? If you:

- Make decisions in an impersonal way, using logical reasoning
- Value justice, fairness
- Enjoy finding the flaws in an argument
- Could be described as reasonable, level-headed

then you prefer
T
Thinking

- Base your decisions on personal values and how your actions affect others
- Value harmony, forgiveness
- Like to please others and point out the best in people
- Could be described as warm, empathetic

then you prefer
F
Feeling

4. How do you prefer to live your outer life? If you:

- Prefer to have matters settled
- Think rules and deadlines should be respected
- Prefer to have detailed, step-by-step instructions
- Make plans, want to know what you're getting into

then you prefer
J
Judging

- Prefer to leave your options open
- See rules and deadlines as flexible
- Like to improvise and make things up as you go
- Are spontaneous, enjoy surprises and new situations

then you prefer
P
Perceiving

An extra-specially-gifted testing man, Lee Cronbach, was at the University of Illinois until 1964. After moving to Stanford, Cronbach tried for a decade to find a "talent platform" on which pedagogy might be based. He called it, aptitude-treatment interaction. "Teach *each* to his or her individual talent." In the Eighties, he and his partner, Dick Snow, gave up saying, "We are sure it's there; we just couldn't find it."[2]

[2] Cronbach, L. J., & Snow, R. E., 1977. *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.

Clearly the contemporary testing and teaching in our schools favor some children more than others. But *fairness* is not always sought. Teachers like myself prefer to teach students like ourselves. Parents *want* discrimination that enhances their children's college and employment opportunities. Grade point averages are here to stay.

Over the years, Joel Spring, an educational philosopher, wrote numerous books about national educational policy. One book was *The Sorting Machine*[3] in which he claimed the primary aim was to identify and favor the children fitting-best the U. S. economic structure.

Societally, is sorting needed? Pedagogically, is tracking needed? Educationally, are tests needed? Standardized tests provide ranks of students. They do not tell what a student knows. They do not tell what a student can do. They do no more than compare students, one to others, on hypothetical talents. Many purport to measure aptitude. They measure neither intellectual function, nor accomplishment nor potential – although sometimes, for diverse groups, they provide scores that correlate with *some* brain function, *some* accomplishment and *some* potential.

There are winners and losers. Equity is not intended. A few test takers get special privilege, many do not. It is easy for an unthinking world to suppose that scoring is neutral—and seldom hurtful. In most places, I think, students are told or have access to their test-score standing. It is easy to conclude that being told over and over, "You are inferior," is damaging. What

[3] Spring, J., 1988. *The Sorting Machine: National Educational Policy Since 1945*. White Plains: Longman Inc.

is “quality of life,” surrounded by people repeatedly lowering expectations of you?

In 1970, my colleague Terry Denny and I, with help from Craig Gjerde and Ben Stake, contracted to evaluate TCITY, the Minneapolis-St.Paul Institute for Talented Youth, a six-week gathering of several hundred high academic high schoolers for intellectual and social exploration. Terry startled the evaluation coterie by asking the question, “What is the effect on the other youth of Minneapolis-St.Paul to have hundreds of their stimulating friends disappear for the summer?” Those chosen. Those not chosen.

Recently, Harvard College has been arguing that, without diversity, "Harvard would lose a great deal of its vitality and intellectual excellence and that the quality of the educational experience offered all students, would suffer." All students. Understanding *is* facilitated by diverse perspectives. Diversity *is* catalyst to a healthy society. We once thought that we could promote equity by carefully measuring talent. We have, today, standardized testing, a major barrier to equity.

Ranking

For my masters thesis, in 1954, I constructed a “quantitative aptitude” test -- a test to identify students, entering graduate school, most likely to do well in statistical studies. I called it the *Quantitative Evaluation Device.* It correlated well with grades in later statistics courses. It competed favorably with the *Doppelt Mathematical Reasoning Test.* Here is one item from the QED:

Four of the five have a property. Mark the one not having it.

(a) length of Joe’s foot
(b) height of the tree
(c) population of Portland

(d) *number of leaves on the tree*
(e) *final score of the game*

One of the characteristics of a test item is its "discrimination index." A good item contributes to the discrimination between high scorers and low scorers. It rewards high scorers and discriminates against low scorers. As a grad student, I bought a few groceries, paid by graduate schools using my standardized test. But, after a while, starting to feel uncomfortable about discrimination, (well, for various reasons, I have trouble remembering.) in about 1970, I took the test off the market.

[(e) is the "right" answer. It was the choice of most high scorers.]

State mandated tests are one problem. Teacher-made tests are a different problem. Most teachers want to give tests, if only to motivate students. They do learn something about what students have learned. Swedish evaluator Ulf Lundgren[4] found that pacing instruction worked well when fitted to performance of students about at the 20th percentile.

Today I am not saying teachers should test less, but they should discriminate less. They are obligated to give grades, and with big classes they need tests.

What is the need to discriminate *among students*? Do we have to compare as much as we do? A great deal of deep thinking involves some kinds of comparison. The same about thinking in general. We talk about living, and about ethics, about caregiving, with attention to function, and problematics,

[4] Lundgren, U. P., 1972. *Frame Factors and the Teaching Process. A contribution to Curriculum theory and theory of teaching*. Stockholm: Almqvist & Wiksell.

and context -- and often we compare with other modes of living and ethics and caregiving. I am persuaded that it is impossible to think, without comparing.

Said one frog to another, “Gosh, you’re beautiful.” Said she, “Compared to what?” We spend a lot of time comparing people. And it regularly means putting some people on pedestals and putting other people down. We compete, partly to appear better than others. I do not suppose we as a people could compete less, or compare less, to think less of what is better and what is inferior. But I wonder if we could be less hurtful. Less hurtful.

It is not hurtful to say that Japanese cars are superior, or that Hershey chocolate is inferior, or that Urbana is a better place to live. But it is hurtful to say that Benjamin is a slow learner. Some stereotyping is inevitable, but we need restraint.

Partly because of national and state standards, we compare students unnecessarily. It is of little help to a youngster to know he or she was at the top again. It regularly hurts students to be shown they were wrong again and again. Standardized tests are norm-referenced, they do not tell what a child has learned, only how many others have more correct answers than he or she. Grades too tell almost nothing about what a student knows.

A teacher needs to know how well a child has performed, and how well classes are progressing, but it doesn’t help a teacher to know rankings in the class. A grade of D for Sarah should not always mean she is below Michelle who quite often gets a C. At least sometimes it should mean how good Sarah just did.

Kids ask each other: “Wadja get?” There is a thirst to know who is better and which is best. Comparison thrives in business, politics, sports and science. And so be it. But comparison thrives also in education. Hurtfully. Comparison of one person against others is simplistic. Especially in mandated courses, if the student has no choice of being there, discrimination is wrong.

The *Universal Declaration of Human Rights* requires respect and protection for each and every person. Comparisons of children are seldom needed for instruction. I recognize parents’ and employers’ desire to know. But aren’t grade point averages, *rankings in class,* a violation of human rights?

Bernadine and I have one great-granddaughter, Sloane. She’s eleven. Sloane will win many prizes. I expect she will live a good life. What should we be doing to make her life better? Better attention to her community could make her life better. She should be living in a more equitable world. She should not have to live where her friends and community are stigmatized by low ranking. We can encourage teachers to use individualized standards over common standards. We can protest against schools that fix students as gifted and having special needs. We can encourage teachers to support the self-respect and equity of students.

Talent flowers with diverse perception, congeals in pursuits of rank.

2019

I prepared a brief Skype statement for a class on Research in Art Education taught by Professor Biljana Fredricksen at Høgskolen i Sørøst-Norge, April 10.

The Diptych of Research

Two paintings or panels hinged together are sometimes called a "diptych." Today, let us put a "diptych of research" on the whiteboard of your mind. Let us hang it there on permanent loan, something you can open up and look at forever. There it is: DIPTYCH OF RESEARCH.

On the left, the panel reads, "Understanding the Question." And on the right, the panel reads, "Portrayal of the Understanding." That is the twin jobs of research: To gain new understanding and to convey that understanding to some other people.

As researchers, we often have a main question, some issue, some shortfall of knowledge, wanting an answer giving us a better understanding. The first panel of our work is to arrange a situation or a panorama of situations so as, sooner or later, to improve our understanding.

The second panel is to compose, to script, to *portray* that improvement of understanding. The researcher is the artist, the teacher, the medium, through which learning can occur. The diptych is a simplistic but slightly elegant representation of doing research. There it is, right there on your whiteboard.

We often think that the first step is to choose some method, a survey or an observation protocol or an experiment to get responses from people. And often some of the understanding will come from people: from our students, our administrators, the subjects of our study. But if we back up a bit, and think, we will realize that what we want first to do is write a plan, a script for action, that will give us new insight into interactions and relationships and interconnectivities, a better understanding.

Suppose we yearn to know better how to teach the complexities of social justice. One issue might be: Will it help to teach males and females somewhat differently? Another might be how to confront the opinion that people get what they deserve. Such issues are the ingredients of our research question.

Our research plan calls for us not just to think of where truth might lie, not just what would be important to observe, not just what instruments might be useful, but how we as researchers might position ourselves to confront the truth, or at least to obtain new insights.

Why do we review the literature of history and philosophy and pervious research, the most relevant research, as a chapter for our dissertation? Because others before us have intentionally, deliberately, placed themselves in situations to confront the truth and gain new insights, and, sometimes, published them. To move beyond the frontier of understanding, we need a better idea of where the frontier is., the review of past research.

To teach the complexities of social justice, we need not only to know the meanings of social justice, but the human perceptions of justice and the aesthetic and systemic ugliness of injustice. What do we need to see? Who do we need to talk to? What do we still need to read? On what grounds should we redefine the meaning of "entitlement?"

From the beginning we will gather some new data, and ever more organized, gather more until we have, maybe too much.

As to the second Panel, even from the beginning we have ideas as to how our experience from data gathering can be conveyed to others. One chapter of the dissertation or section of the report may be devoted to describing: not how the data gathering was supposed to go but what actually happened. The meaning that readers will give to your insights will be influenced by how they see your involvement in data gathering.

Of course they will interpret some of your understandings differently. Not a bad thing. You should take credit for promoting interpretations beyond your own. To earn that credit, you should have pondered different ways of writing the scripts for sharing your understandings.

Suppose you had asked certain students if they found social injustice in their own classes. And one of them said,

> *Yes, one of the things we learn in all classes is that we will be judged, that we will be graded. Being graded is part of the culture, which we need to know. But being graded is also a treatment. And it is an unfair treatment to be regularly told we are deficient in these classes, because we take it to mean that we are seen as deficient intellectually, that we are second class persons.*

How will you use such a (hypothetical) statement in your portrayal of social justice being learned? Your data need to aggregate to findings. You need to have a certain consistency in your final writing. But you also need to convey the inconsistency and incompleteness of your study.

Your research report will not be elegant because it finds some perfect solution. That won't happen. It will be elegant because you put yourself in situations to learn and because you wrote a good script that conveys to others what you learned.

Many social scientists have urged making the research as impersonal as possible, to make it objective. The diptych on your mental white board says almost the opposite. The research has to pass through the triptych's best arrangements the researcher can make. Even the most objective research has to be subjective.

It can be done with just one set of hinges.

2019

After years without one, in 2018 we got a cat, something perhaps to ease Bernadine's dementia.

Coping with Voids

In 1972, I knew the importance of multiple views. I knew the difference between direct and indirect measurement. I knew the difference between a cat and the experience of owning a cat.

In 1972 I was a measurements man, and I still am, here in 2019. Back then I boasted that, given tools and skill, I could measure almost anything. I could measure a cat and I could measure the experience of getting a cat. I am less sure today.

By measuring a cat, I don't mean just getting its length and reaction time and appetite. I mean representing a particular cat in the many ways that someone wants it represented. That's one way I would take my measure of a measurements person.

From time to time, a measurements person has a client or maybe an audience, maybe an audience of stakeholders, people having a stake in the evaluand. Let's call it *evaluand* because I don't have a better word for the thing being measured.

A good measurements person ("mensurer," sometimes so-called) is a service provider, providing information intended to be helpful. It could be for a good cause or a bad cause. An ethical measurements person tries to serve good causes. A mensurer doesn't stop being one just because he or she buys into unworthy causes.

A good measurements person provides information that is accurate. Sometimes a client needs highly accurate information, sometimes not. The good measurer tries to find out both what the stakeholders want and what they may need.

It is easy for a measurements person to be wrong about what stakeholders need, especially if they strongly want something else. It is easy to be wrong by having a business model for measuring only what you think they need.

Of course, the representation can be poor because it isn't what the clients want or because it isn't what they need. And clients (and even one client) don't always agree. For some, the representation can be quite accurate but not really good.

Making a highly accurate representation of an evaluand often is quite difficult. In representing much social activity, accuracy probably means that some characteristics have been underplayed. Accuracy and goodness may be like oil and water.

Owning a cat can be a complex experience. In various ways one can tell the story of that experience. Views differ and change. In fact, the story may change the experience.

How one measures inevitably changes with fluctuations of the evaluand, with the context, the audience, and the personality of the mensurer. Evaluands differ as to materiality. After you represent a cat, the cat is still there. After you represent an experience, the experience is gone.

The cat is material. The experience is immaterial. Sometimes the measurements person fails to realize how immaterial the experience was.

Good measuring is not just a matter of aggregating the experiences, particularly those of a number of people, a task that has its own dangers. With each experience there was a happening, a movement, a fulfillment, a shortfall. Now gone.

Like the tree falling unheard, if without personal awareness or after-effects, the experience did not happen. Measurements people run risks making representations, telling stories, based on experience. Coping with voids.

And yet, there is no lode so rich in data as experience. Ultimately, measurements of the cat tell so little. The experience of owning a cat can be material for reporting on matters of life. Even with little accuracy, a measurements person's representation can reach further toward goodness, toward being of service.

2019

A student asked me if case study could be useful for clarifying the term, "legislative governance." I wrote him more or less the following.

The Meaning of Meaning

Let me say first that names, labels, and, yes, all words, are inevitably imperfect representations of the phenomena referred to. There are no true, or even completely shared, definitions. Rene Magritte painted a pipe and named it, "This is not a pipe." There is only a limited correspondence between a thing and its representation. A meaning is meaningful but is not the thing itself.

Definitions can be agreed upon through some expression of unanimity but meanings can only be meaningful through human thinking. The terms, even the most abstract, are rooted in personal experience. And inevitably experience will vary from person to person. We can accomplish a lot with language because we share many meanings and experiences, but there will always be differences in meaning. Often it is useful to work to reduce the differences and that is what you aim to do with the meanings of modifiers of "governance."

We expect that an important function such as governance will be conditioned or refined by the use of modifiers. "Corporate governance" speaks of the setting or ambiance. "Good governance" is a term useful for political advocacy. "Evidence-based governance" suggests reliance on disciplined use of accounting. Modifiers steer the meaning toward both fewer and more purposes.

The modifiers may be used to clarify the meaning, but they often are used to increase the standing of the user. It might be useful to have an agreed-upon lexicon, but many would not agree to abide by it when allusion fits their purpose. One purpose of research is to show the diversity of meanings to be encountered and even perhaps the costs and benefits of ambiguity.

Often it is more or less too much to ask people to define their terms. We should get what we can from them, but knowing what those individuals themselves mean is an advanced knowledge. Regularly people know more than they can tell us. Partly we get what they want us to think they mean. Still, they just can't say all that they have been thinking.

James Reston, the *New York Times* journalist, once asked, "How can I know what I think 'til I read what I write?" That is an elegant wisdom. No one writes exactly what they think, oftgen not even close.

And more, we do not think in words. We think in awarenesses, in experience. We are aware of happenings, real or imagined, about us. Some of our thinking is the experience of translating, in our minds or in writing, experiences into words. The words are a product of our thinking. We can think about words, but not in words.

What we express, we won't necessarily express it the same next time. Case study can be helpful because, as I would have it, it calls for looking again and again and again, and seeing some of the consistencies pile up.

I don't propose doing a case study on concepts or conditions or phenomena. What is the case to be studied? I usually want the case to be personal, perhaps an organization of human beings, or a happening that humans are experiencing. An event. I expect the primary interest in doing the case study is to gain understanding of that particular case. That doesn't really include studying how a concept such as governance is variously used.

But, whatever I say, I don't own the method of case study. Others may decide to use case study to investigate a certain happening, such as the workings of your Governance and Management Legislative Network, to reveal the meanings, then and there, that are being given to the various concepts of governance, particularly “legislative governance.” They may find that it is primarily used as a weapon to counter administrative governance when the executors of policy violate the avowed intent of the legislative body. Or some such. Their meaning may pertain to a certain collectivity in a certain place and in a certain way at a certain time and additionally limited. In other words, the case study may help understand the workings of a body, including the language used. But of course it wouldn’t claim to offer true meanings.

I once had a doctoral student, Judy Dawson, who wondered if, how, and when, curriculum specialists, working in state government, influenced the achievement testing mandated by the state. For a year, she sat in their offices and conferences, spying. Not furtively. She read and overheard the workings of the two departments. She concluded that the curriculum folks were only paid attention to by the psychometricians when they talked in psychometric language -- which did not always lend itself to their concerns. So it was a case study of the their work, a happening, the interaction, the executive interaction at given places and times, with the issue of communication and lack of communication prominent.

Several years ago I observed an Uppsala researcher (Judit Novak) studying the contemporary trending of “juridification,” the drift of administrative governance toward legislative and judicial governance, the increase of legal trappings and constraints. It is worthwhile looking into her work.

One could make the mistake of focusing too much on any term, such as “legislative governance” or “getting a cat.” The concept will be broader than the name. There are other ways to reference the concept. Various users will have various ways of imaging it. Their stories will seldom be constrained by definitions.

2020

When Merel and I started working together, thinking together, living together, we acknowledged that we both were writing about doing program evaluation humanistically. Our thoughts went something like:

Making Research More Humanistic

If there is a Humanism, and if all humans have a duty to contribute to the integrity and preservation of that spirituality, then it befalls professional researchers to guide disciplined inquiry toward its support.

Even if there were not a system of thought or action presently based on the nature and ideals of humankind, it befalls professional researchers to raise the question of how their special fields of inquiry are elevated and obstructed by human efforts.

Even if there, at this time, were no abiding human intuition or agency for the betterment of humankind, it befalls professional researchers to use part of their inquiry for identifying opportunities to enhance the well-being of the human race. Are there any fields of research that are excused from this duty?

If these are among the collective duties of the research community, what are the ways, or at least what might be a first next way, of changing the research they do?

A first way is not to impose upon the sanctity of freedom of choice of life work as to how much an activist social-reformer a researcher should become. It is not necessary for any human to become a humanist, a vocal advocate, in order to enhance the well-being of fellow humans.

If such an individual duty exits, it is the obligation of every professional association to dedicate some of its efforts to the integrity and preservation of Humanism -- for it is this concept of duty for thought and action without which the association can form and remain independent of political and ideological control.

There may be others, but a candidate for first way is the obligation of individuals and collectives to support Humanism by raising the following question in all inquiry spaces entered, the question of "Is this a space in which a sense of justice and fairness and caring is to be found?"

A mere mention of the question will not assure movement toward justice, fairness and caring, but engagement in the thought required by such questioning should motivate subsequent steps in fulfillment of that duty.

Such steps will still be far from adequate to create or sustain a Humanism, but they should be steps away from the contemporary status of collective research which tolerates and even promotes productivity, efficiency and financial gain over the betterment of the human race.

2021

From First Grade on, starting with Miss Maston, I saw a lot of Education. Some 22 reflections:

At 90, In the Rear-View Mirror

1. The education system is very complex. It is not just an aggregation of all the individual schools. We need to be acquainted with individual schools but we won't understand the district or national schools just by being acquainted with lots of individual schools. Nor even from any survey or visitation to all the schools.

2. The schools are one of the largest national expenditures, so they figure into people's thoughts of the fairness of wealth. Still, rich and poor people agree pretty much alike: "If the schools are weak, we shouldn't give them more money to waste."

3. There is some truth to the conspiracy theory that people in power, preferring to spend less of the national wealth on schools, have looked and found ways of making the schools look worse than they are.

4. Schools themselves are environments much worse than the lives that we want our children to be living. Schools themselves, as environments, are better than the lives a lot of children are living.

5. One cannot know the quality of schools by testing their students or by listening to school leaders. Or by any of the indicators economists have been devising.

6. Standardized achievement tests basically rank students as to long-lasting aptitudes and to intellectual privileges experienced. Aptitude is highly correlated with scholastic

learning, and more easily measured, which makes the test scores weak measures of disciplined learning.

7. Tests are culturally biased, but that is not seen problematic by most parents.

8. Teachers need to know their students individually. There are no good reasons for comparing students, particularly as to aptitude, achievement and decorum, individually or in populations. But parents, politicians, teachers and the society have appetites for comparisons.

9. Standardization of pedagogy and curricula makes it easier to operate the schools, but standardization exacerbates discrimination and denies the best possible education for many students. "Minimum standards" have been poorly reasoned and are divisive, all who try pass, schools look good, and all are robbed of opportunities to be educated.

10. People do not agree on priorities for schooling. For the very purpose of improving democracy in our countries, we can argue against popular opinion, but we should yield to it in considerable degree in organizing our schools.

11. For most parents it is highly important that their children receive credentials as provided under the existing curricular, testing, attendance and comportment rules, more important than being widely experienced and broadly educated. Credentials limit opportunity.

12. People tend to oppose tax support for schools that are unlike what they are familiar with.

13. Many people have trouble seeing that a multi-cultural society has merit that a homogeneous society does not.

14. If we support democracy, we can argue for diversity but we should yield a lot to public desire for assimilation and homogeneity.

15. Parents should have choices as to the overall content and styles of teaching they want, with occasional opportunity to

change the mix. They rely on custom more than advertised benefit. So do most teachers and administrators.

16. Teachers colleges, staff development, and unions have little influence on content and style of classroom activity. Each could be better, but were they perfect, they would be widely opposed

17. Current practice is a blend of biases.

18. The political system slowly, grindingly, has had an effect on teaching quality, mostly negatively. It has limited how a teacher can diversify, enrich and individualize teaching

19. School administrators in the USA and elsewhere have little expertise in Education, but expertise in organization and public relations. Most know the compliance within their schools, but not the quality.

20. People should have more information including more critical information than they are getting as to different evaluations of their schools. What is really bad? Quiet critical information is needed even though it contributes to lack of support and funding for the schools.

21. Communities and neighborhoods should have the schools they want, neither state nor federal aspirations should rate higher. Teachers, researchers, philosophers and others who have clearer expression of what education could be, should strive to help the people have the schools they want, however inferior, in some ways, they will be.

22. Home schooling, alternative schools, charter schools and private schools should be supported to help give parents and communities the range and thrust of choices they want.

23. Private and semi-private schooling undercut the public school system, especially by robbing the considerable contributions to teaching and learning that public and non-public students then no longer provide each other.

2022

The essays for this book were complete, ready for the printer. But then a family concern came up. And I wrote this:

A Reconciling

The Indian School, at what was to become the village of Genoa, Nebraska, was a one-room school until 1884 when gradually replaced by the Indian Reservation Agency and a multi-structure brick, boarding school on a mile-square site. The school closed in 1934. When I was a boy, it was a pride in Genoa. Now the school is said by some to have been a travesty.

Over a hundred Indian young people apparently died there across those years. There is no doubt the sanitation was poor. Tuberculosis was reported the main cause of death, but suicide and shooting also were told.

A reconciliation effort exists today. Efforts are under way to find a cemetery and more bodies. With enrollments sometimes over 200, a year might have taken several kids -- in 50 years, maybe, more than a hundred. Some deaths surely could have been prevented. We can suppose that most would not have died had they remained with their parents.

Not enough records were kept. Not enough is known for an ethnography of operation the school, not even for the questions the reconciliators would like to answer. Some shocking family stories were told and a few former students wrote exposing accounts. Happy stories too, but not ruling out a picture of mistreatment, a long perpetuating isolation, a curriculum too little accommodating the life the children looked forward to. On the other hand, the current museum guide presents an ordinary picture of training, a picture of reason and safety. The pictures are at odds.

For over 200 years prior to the creation of the boarding schools, the policy of the United States was to dispossess native Americans of their lands and lifestyle. It included genocide, slowly. After a more sudden genocide, the Civil War, the question pushed forward as to what to do with Indian children. Army Colonel Richard Pratt proposed they go to school, learn English, become vocationally skilled, and assimilate as immigrants into the national population of 50 million. Forced "Americanization." "Colonialization." But also, according to historian Brenda Child,[1] faced with poverty, some Indian families were happy to find institutional food and care for their children. Still, separation was cruel.

As in many places, the original small Genoa school was started a few miles from a Loup River settlement of the Pawnee tribe, in flat, open country. Not long after 1879,[2] when the federal government relocated these pastoral people to Oklahoma, opening their farmland for White settlement, the Genoa Indian school was greatly enlarged, taking in children from tribes across the country.

As scouts marked the trail for the Pawnee to walk their way through Kansas to the new reservation in Oklahoma, as a cook helper, they took along a nineteen-year-old boy, Joe Coffin, riding a pony. Later he was my grandfather. Their trail passed not far from where I was born in 1927.

Joe Coffin was born in 1858 at Chatham, NY, losing Susan Coleman, his mother, in childbirth. His father, a farmer, apple grower, William Coffin, grieving, left Joe with an aunt, and took a job, 1400 miles away, to teach agriculture at the Genoa

[1] Child, B. J., 1998. *Boarding school seasons: American Indian families, 1900–1940*. Lincoln: University of Nebraska Press. p. 55.

[2] I am not certain of my dates.

Indian School. A year later, maybe more, William returned to Chatham to marry Susan Robinson and bring her and his three children to Genoa. He continued as teacher and farmer, dying in 1893. His gravestone at the Quaker cemetery said, "A Friend of All."

Joe attended Peru Academy on the banks of the Missouri River, homesteaded in the sandhills two counties north of Genoa, started a short-lived well-digging business, and married another Quaker, Maggie Foster. Joe had no formal connection to the Genoa school, but somehow accumulated a collection of Indian artifacts. Long after, not so long before the Indian School closed, Joe turned over his collection to the Nebraska State Historical Society. I added a wampum belt he had given my mother. We didn't learn the circumstances of Grandpa's acquisitions, or the reason some folks had called him "Boy Chief" and later "Pawnee Joe." Perhaps he chose the latter for his costumed dance in Denver Dick's Wild West Show.

My mother and Aunt Jennie Fleagle loved Grandpa Joe. Jennie wrote me in 1983: "He was a kind, lovable person, quick on the trigger but held no grudge -- while she (Grandma) was so different, gentle, loving. We children never! never! never! heard our parents argue or quarrel. There was always peace." I was told he, coming home from the city, had surprised Maggie with a sewing machine, an item new on the market. I asked if he had known what kind she wanted. My mother had said that then it wouldn't have been a surprise.

I spent a good number of underage 1930s summers in Genoa, living with my cousins and their mother, Mamie Hickey, a newspaperwoman. Perhaps there had been stories around town about teacher cruelty. The family didn't tell them in my hearing. Lots of talk, none I can remember about trouble at the Indian School. Nothing of maltreatment. Nothing of dereliction. Today: a search for a cemetery, presuming more children's bodies to be found.

I would hike with the neighbor boys to swim in the gravel pit near previous Pawnee places. The boys filled my ears with rudeness. They bullied me, popped my blisters. I can't recall any stories about Indian students having to wear uniforms, have cut hair, speak English, never to go home for vacation. I know now it happened, but neither did I hear of inhuman punishment, exposure to cruelty, ill-kept food, over-exposure. Perhaps it was covered up. What to think? I cannot disparage my grandparents if only because the world has atrocities -- Holocaust and popular support of slavery.

What I have taught my students, and should apply here, is what historian Brenda Child observed:[3]

> *...there are different eras in the history of American Indian education. And so what Native people who attended a government school might have experienced in 1879, when there were still Indian wars being fought in the United States, was quite different than what [an American Indian] student in the 1930s experienced when people in government were saying, "Well, Native people shouldn't have to give up their languages or their cultures." That's a very different period. I don't think that students who attended boarding schools experienced the same thing decade after decade.*

Of course there were deaths. Of course there was a cemetery.

[3] Child, B. J., 1998. *Boarding school seasons: American Indian families, 1900–1940*. Lincoln: University of Nebraska Press. p. 55.

What is the mark by which we might know the meaning of what happened long ago? History tries. Memories fade. Diaries embellish. Anger roils. Vows and laws and religions and rule books -- no matter when and how constituted -- tell us little of how to prioritize a present-day boundary that was violated back then. We have long counted on impulse, indignation, embarrassment, and trusted advice. Those worked imperfectly then; no better now. A shrill voice of modernity urges us to feel that the past was cruel, that: old righteousness was essentially discrimination. And, an ever-rising "virtue of care" helps makes it easy to overstate the wrongs of times past.

Disinformation has come of age. Faster than we can put together words to mark how we should live, how we should care, we are caught up in media storms that bind us close to narrow culture, with norms, with keepers of our trust. Storms of video and print, education and speculation, rife with overstatement. We think too little for ourselves. We cannot find a true measure of hurt -- all we truly have is the hurt we feel.

Among social phenomena, there is no better palliative than reconciliation. Moral justice. A sanctity. Reconciliation acknowledges the tortures of the past. It isolates and tries to integrate combatant forces. It identifies aggressors and victims, but more important, it appeals to forgiveness, and promises a fulfilling expression of respect. It loses its sanctity when it is hell-bent on seeking inventory of evils of the past.

Nebraska Commission on Indian Affairs Executive Director Judi gaiashkibos was quoted[4] as saying, "… this boarding school is not like the ritzy, fancy boarding schools out

[4] https://nebraskapublicmedia.org/en/news/news-articles/archivalreview-brings-known-genoa-indian-school-death-toll-to-59/

east where their being was enhanced. Ours was diminished. They were destroying all these children."

My great grandpa taught the Indian children agriculture, a respectable vocational subject. As is still thought appropriate -- consistent with having the boys do what their fellow American prairie boys were doing -- he probably assigned them work-experience (working on nearby farms, including his own). Perhaps William diminished them.

Co-Director of the Reconciliation Project is Margaret Jacobs. Speaking of the deaths of students, she was quoted,[5] saying,

"It kind of surprises me that we're not finding them in the government records. It kind of suggests to me -- and this has been confirmed by some colleagues of mine who work on this issue -- that the government really tried to cover up these deaths."

Speaking further of deaths, gaiashkibos added,[6]

"Will we ever know the truth? How many? But I think until we bring closure to this, we can't move forward and heal."

In any faculty of 500 teachers, there will be rascals, even criminals. William? I have no evidence, but I can't believe it of him. A few teachers leave traces of cruelty and negligence. And discriminatory blessings and false pride. Sometimes the parents support it. I have heard: "But it's the right thing for those kids."

[5] Ibid.

[6] Ibid.

I am talking about disparagements, leaving lasting hostilities, painting in White superiority.

Bodily hurt and personality injury are seldom the effects of a school. I think it unlikely that the Indian boarding schools were Andersonville prisons. I think it likely that most of the coming and going 500 staff members of the Genoa school thought a stern and military tenor were right for the children. Perhaps also for the children of their own families. Almost two centuries later, we fail to care well enough for the children, but it is as much in excessive tolerance and shared discrimination that we fail. Times change.

The tragedy of boarding school treatment of Indian children was government policy, with citizen acquiescence. Little of it, I think, was school practice: not ambiance of the classroom, nor food at table, nor sanitation. But, tragic was the separation of families. Robbing parents of love and care and responsibility for their own children was a crime as great as stealing their prairies, rivers and buffalo. Robbing brothers and sisters, grandparents. President Grant thought that staffing the schools with Quakers would compensate -- but not even close. Nor would ritzy boarding schools in the East.

The boarding schools were ordered "to take the Indian out of the child." It was thought, and sometimes still is, that there is a personality syndrome characteristically native American. Differences in the many tribes and diversities within families belie that notion. Americanization should not be homogenization. Separation is not therapy. To mandate separation was a travesty.

And is. William was not a suitable mother for Joe. Nor was he a suitable father for any Indian child. I wrote a poem to express my feeling:

We stole their gravel pit.
Then tried to pay with boarding schools.
It didn't work. Yes, a travesty.

Not the haircuts, nor tuberculosis,
but the family separation.
But we tried to care. It wasn’t an ignoble try.
Something we thought we knew how,
but didn’t.
Now we’re counting graves.
It won’t work.

Reconciliation isn’t
counting every last grave.
It’s trying harder.
It’s stop making guns.
It’s welcoming refugees.
It’s stop comparing children.
These reconciliations won’t work either
but it would be trying
to pare down our enduring travesties.

2022

If you take it a little further, "everyone is different" becomes" everyone is different in every way." My brother Don and I are not just different, we are not the same in any way. Yes, male, same mother, and we wear the same shoe size, but I am trying to be serious about this. There's nothing about us that is the same. I probably have said, "Yeah, Don, he's one of a kind." But that implies there is a bunch of people of his kind, that there are many quite like him. There are not. He is not one of a kind, he is not one of a herd. He is a kind of one. That is to say, of his category, there is only one, and he's it. The next closest one, and it may be me, is one of a different kind.

Some Kinds of One

Sooner or later, we all hope to acknowledge those who have taken care of us and those who have influenced us. Along the way in this memoir, I have mentioned quite a few of mine. Too seldom have I sufficiently detailed the nature of that care and influence, much less measured it, although that is a major purpose of this writing. You may remember it started as, "Through a Measurement Darkly." Those caregivers were there at critical moments of my life and left a well-after-birth-mark on me.

Back in Nebraska, skipping my father Earl and mother Nelle, (she preferred Gertrude but that was never her legal name), perhaps the most influential person was John Leach. He had chronic headaches and almost daily would come to the Stake Drug Store where I would fix him a Bromo-Selzer. He ran the Palm Café and Palm Theatre and let me sell popcorn to those seated before the feature started. Our banker Heini Gramann had John keep track of the jobless men available to help with farming and trucking when needed. Some sat on the bench outside the Palm Café. I came to think of him as an unofficial caretaker for the community. John Lamason, the Adams school superintendent, probably came next in influencing my mind as to caretaking.

When I graduated from high school and went 35 miles north to the University of Nebraska, I lived at the Brown Palace, a Men's Co-op, hanging out mostly with Walt Sehnert and Chris Buethe — "Stuethnerts," we called ourselves. I took a writing course taught by Oliver Wright. Given opportunity, his eye'd catch the dangling participle. Stuck in a snowdrift, hitchhiking with Dr. Swartwood, Prof. Wright let me take the final exam late.

Months later, on the Fourth of July, Don Schneider, Lee Dyer and I, avoided the army by joining the Navy. There I met Gary Joselyn who taught me how to "throw shoes." Our motto later became: "Age and treachery will win over youth and skill." Sixty years later, Gary's wife Yleen, part hummingbird, larger part caregiver, confided, "I'll die broken-hearted." Think that meant her To-Do list was still too long.

Picking up Stakes in 1954, we moved to Princeton. I plunged into psychology, would probably not have survived had not Laury Gulick prepped me for the qualifying exam. I had met Warren Finley at a baseball game in Lincoln. He was one of the vice presidents of the Educational Testing Service and he had encouraged me to go for a Princeton doctorate. At the time, the American cities were finally, gradually, integrating their schools, and Atlanta invited Warren to help them assemble data to persuade people that, intellectually and academically, the white and black kids considerably overlapped. When I needed eight classrooms of children to do my doctoral research, Warren invited me to do my testing in Atlanta.

My Princeton mentor was Harold Gulliksen, a renowned writer of psychometric theory. He was special, unique, a kind of one. His doctoral herd at the time was small but well endowed, including: Fred Kling, not long out of seminary, who invented a great game, Information Hex; South African Bruce Faulds, who urged me to stop writing racial quips in my Christmas letters, irritating the censors; Hal Schiffman who chided me for choosing my clothes the night before; Don Thomas, who at least once unconsciously mimicked Gullickson's murmuring, "la-le-da-da," on finishing silently reading a paragraph; Carl Helm, who after a car crash said he'd longed for a face scar; Ron Weitzman, whose fish for his doctoral research died in a Christmas vacation boiling; and me. At least the others, a kind of one.

After Korea in 1953, rental prices returned me to Lincoln to study educational measurement, $35 a month rent at the boarded-up air force base. Chuck Neidt, a genial, caring chair of ed psych chair, steered me to a newly arrived graduate student to advise. Tom had little interest in testing, but measuring football-linemen skills seemed to him worth doing. His name was Tom Osborne.

We were Cornhuskers but Bernadine and I moved our four kids (Jake in utero) and kittens to Urbana. In earlier pages, I have identified Tom Hastings, Jack Easley and Terry Denny as long-time down-the-hall folks to be counted on. Gordon Hoke was nearby, as were still two more Terries, Elofson and Souchet, and, of course, Barry McGaw, Craig Gjerde, Paul Theobald, Linda Mabry and Shameem Rakha. In the Dean's Office there were: Rupert Evans, Mike Atkin and David Pearson. I'm leaving out lots. Better to list all the CIRCE people. I'll do that after I close this last essay.

I should make special mention of the relatives who helped one time and another and another: Don and Nancy Stake of Sunnyvale, Dick Madden of San Diego, Claryce Evans of Harvard, Mamie Hickey and Lydia Cochran of Genoa, Edna

Kuster of Lincoln, Alan Lemke of Bloomfield, Elba Saavedra of Albuquerque.

Great help from abroad were Ulf Lundgren of Sweden, Pepe Aröstegui and Fátima Cruz in Spain, Giordana Rabitti of Italy, Ami Maiga of Mali, Stephen Kemmis of Victoria, Oli Próppe of Iceland, Edith Cisneros of Yucatan, Penha Tres of Brazil, David Metzer of Israel, Bob Louisell of Duluth, Eva Baker of L.A. Margaret Fiore of N.Y. And many more.

Most of those who should be offended by my failure to mention their names and heroic rescues, won't be, because I didn't get around to writing this book soon enough. Truth is, and most of you know it, the cares and nudges were minute by minute lifelong, beyond those of family and co-writers and co-everythings. And each and every one: a kind of one.

Here's a last thought: Psychometrics "officially" belongs to Individual Psychology. Psychometrics is the measuring of individual intellects. But psychometric scaling and testing, I found out, are the technologies of grouping. In this applied science, we measure primarily to put people into groups, classes, ranks, tiers. We coddle rounding errors and flaunt classification. Psychometrics is the science of stereotyping. Psychometrics includes qualitative methods, particularly case study, but not seen as a refinement of quantitative methods. It's a grand embellishment, as I have said in these essays.   One of a kind is a classification. A kind of one is recognition of the mind-stretching uniqueness of each one of us. Each of us: a kind of one.

CIRCE People, years

Visiting Scholars in italics

Illinois Statewide Testing Office

Tom Hastings	45-80
Milosh Muntyan	45-46
Don Thomann	46-48
Evelyn Ropp	47
Frances Aronoff	47-48
Lee Cronbach	48-63
Winona Kott	48-49
Henry Shearouse	48-49
Wilbur Oldham	48-49
Earl Foreman	48-50
Myron Lieberman	49
James Dyke	49-51
Lois Williamson	49-76
Albert Eckert	50
Dora Damrin	50-51
Ed Wahl	50-51
Russ Kropp	50-52
Gabe Della-Piana	51-52
Joseph Edelen	51-52
Beatrice Aaron	51-52
John Hunt	52-53
Jack Merwin	53-54, 73
Dick Spencer	53-54
Mohammed Quereshi	54
Norval Pielstick	55
Nancy Whitman	55-57
Khossrow Mohandessi	54-56
Promila Gupta	57-60
Mike Atkin	58-79
Sefik Uysal	60-61
Gerald Larson	60-62
Jack Easley	60-91
Dave Krathwohl	61-62
Phil Runkel	62-63

(CIRCE was created, 1963)

Gary Marco	61-63
Clement Adelman	63 & 82
Husnu Arici	61
Kazutaka Furuhata	61-64
John Bianchini	62-65
Hiroshi Ikeda	62-65
John Tomczyk	63
Doug McKie	61-64
Larry Weber	63-67
Bob Stake	63-17
Leo Hicks	63-65
Suleyman Ozoglu	64
John Pyper	64
Hiroshi Ikeda	64
Russell Zwoyer	64
Peter Taylor	64-66
Akihiro Yoshida	64-67
Tom Maguire	64-67
Gene Glass	65-67
Sid Dunn	*66*
John Ahlenius	66
Tom Anderson	66
Aletta Ellico	66
Jay Millman	*66-67*
Jerry Faust	66-67
Tom Bligh	66-68
Don Bosshart	66-68
John Paraskevopolis	-68

Duncan McQuarrie	66-70
Norman Bowers	67
Jean Bowen	67
Michael Ellis	67
Marc Gold	67
Brad Hastings	67
Hank Slotnick	67-69
Ira Langston	67-70
Jim Leach	67
Margie Pjojian	67-70
Gary Storm	67-70
Joe Steele	67-71
Jim Wardrop	67-75
Mary Anne Bunda	68-71
Bill Flottman	68
Graeme Watts	68
Phil Wickersham	68
J. Q. Adams	68
Terry Auger	68
Don Cunningham	68
Trey Coleman	68
Sally Pancrazio	68
Sandra Savignon	68
Dennis Gooler	68-70
Alan Koller	68-70
Ed Kelly	68-71
Terry Denny	68-17
A Grotelueschen	68-81
Gordon Hoke	68-17
Sid Dunn	69
Richard Thornes	69
Dan Stuempfig	69
Tom Slotnick	69
Bob Louisell	69
Arlen Gullickson	*69*
Goeff Driver	69
Hubert Dyasi	69
Jirawat Wongswaddiwat	69
Ernie House	69-89
Steve LaPan	69-70
Tom Kerins	69-70
Harry Robinson	69-70
Pam Rubovits	69-70
Don Bosshart	69-70
Gary Storm	69-70
Michael Plog	69-71
Clencie Cotton	69-71
Sue German	69-71
Larry Rosenkoetter	69-71
Norman Stenzel	69-71
Margie Pjojian	69-71
Doug Sjogren	69-71
Barry McGaw	69-72
Joyce Riley	69-73
Bernadine Stake	69-81
Gary Joselyn	70
Merl Wahlstrom	70
Virginia Gonsalves	70
Shirley Kessler	70, 78-79
Christine George	70-71
Ludwig Nemeth	70-71
Paul Elliott	70-72
Heather Sharman	70-73
Terry Elofson	70-73
Bob Wolf	70-74
Barbara Schneider	70
Gabriele Lakomski	70
Martin Maehr	70
Larry Ingvarson.	*70*
David Addison	70
Larry Cross	70
Jo Friedman	70
Jerry Gage	*70*

Marjorie Carss	70
Carol Petrie	70
Jean Chen	71
Marjorie Lee	71
Ethel Mincey	71
Janet Ostrand	71
Jerome Seelig	71
Craig Gjerde	71-73
Mara Stolurow	71-72
Valerian Summerville	72
Valerie Soderstrom	72
Irene Dowdy	71-73
Nick Smith	71-74
Diane Essex	71-74
Marty Hotvedt	71-74
John Long	71-76
Deb Rugg	71-86
Harry Broudy	*-81*
Sue Funnel	72
Eileen Broderick	72
Christine Davis.	*72*
Don Hogben	*72, 78*
Jonathan Block	72
Pat Walsh	72
Royce Sadler	*72*
Kathy Brophy	72-73
Tom Anderson	72-73
Tom Lawson	72-73
Tom Grayson	72, 90, 14
Stephen Kemmis	72-76
Marc Tabackman	72-73
Helen Simons	*72, 76, 79*
Michael Bakalis	73
Marge Steinkamp	73
Tamara Sihon	73
Arthur Boynton	73
Barbara Francis	73
Joseph Green	73
Gillian Ortony	73
Floyd Pennington	73
Helen Rose	73
Steve Rugg	73
Eric Weir	73
Rick Amado	73
Linda Bock	73
Jon Bowermaster	73
Mary Jensen	73
Barbara Schonborn	73-74
Rich Sudweeks	73-76
Darrell Caulley	73-78
Linda Baker	74-75
Judy Dawson	73-79
Oli Proppe	74-75, 78
Monika Meinhold	*74-75*
Barbara Kremer	74-76
Arieh Lewy	*75*
Ruth Knox	75
Hans Brugelmann	*75*
Sue McBurney	*75*
Nick Hastings	75
Joe O’Shea	75
Linda Perkins	75
Al Phelps	75
Frances Anderson	75-76
Mel Hall	75-78
Diane Reinhard	75
Gail Parks (L Armstrong)	75
David Balk	75
Derek Taylor	75-77
Tom Watkins	75-77
Ulf Lundgren	75-78
Pat Templin	75-78

Georgia Cuddeback	76
David Hamilton	76
Rob Walker	76-77
Mary Lee Smith	*77*
Turee Olson	76
Harry Broudy	75-81
Del Harnisch	76-86
David Shramm	*76-77*
Gary Laumann	76-77
Elaine Stamm.	76-77
Peg Steffensen	76-80
Norman Bowman	76-78
Beth Dawson	75-77
Chuck Secolsky	76, 13-17
LuAnn Smith	76
Rich Zollinger	76
Barry MacDonald	*76,78*
Dennis Fisher	77
Betsy Hutchins	77-78
Gary Laumann	77
Kip Anastasiou	*77*
Jo Day	77-78
Jennifer McCreadie	77-80
Kathryn Au	78
Jo Friedman	78
Jeri Willen	*78*
Linda Bolayer	*78*
Vilma Bravo	78
Joan Sextro	78-79
Graham Maxwell.	78-79
Claire Brown	78-81
Peter Fensham	*78*
Bob Carlson	*78*
John Nisbet	*79*
Gera Millar	79
Cal Zinkel	79
Richard Smith	79
Oli Proppé	79-80
Jane Visser	79-80
Gabriele Lakomski	79-80
Deborah Trumbull	79-83
Donna Koenig	78-80
Richard Pring	*82*
Joyce Jones	80-82
Marli Andre	81
Marilyn Munski	81
Mary Lynn Boscardin	81
Michael Eraut	*81*
Sonia Williams	81
Eileen Broderick	81
Sandra Mathison	81
Hallie Preskill	81-82
Janice Hand	81-82
James Pearsol	81-83
Sue Hill	81-82
Colleen Frost	81-85
Mary Krick	81-82
Robin McTaggart	82-83
Connie DellaPiana	83-84
Noreen Michael	83-84
Rosalie Torres	83-84
Corrine Glesne	84
Jan Merchen	84
Penha Tres	84-86
Steve Lichtenstein	85
Mellen Kennedy	85
Bob Linn	85
Jim Raths	85
Leigh Little	85-90
David Metzer	86-89
Liora Bresler	87-89

Paul Theobald	87
Linda Mabry	87-95
Buddy Peshkin	88
Jacquie Burnett	88
Lou Smith	88
Rudy Serano	88
Rob Walker	88-89
Berit Askling	*88*
Aminata Maiga	88
Christi Bergin	88-89
Jonathan Block	88-90
Joe O'Shea	88-90
Giordana Rabitti	89-91
Phil Holmes-Smith	89
Craig Russon	90
Susan Bruce	90-91
Ruth Whitelaw	92
Mike Harmon	92-94
Carmilva Flores	92-96
Mindy Miron	93
Chris Migotsky	93-96
Colleen Medley	94
Mikka Whiteaker	95
Tresa Dunbar	95
Terry Souchet	95-98
Edith Cisneros	95-99
Chris Dunbar	96
Iduina Chaves	96
Carol Mills	96
Stephen Guynn	96
Rita Davis	96-98
Nicole Roberts	96
Kayleen Irizarry	97
Kathryn Sloane	97
Gary DePaul	97
Marya Burke	98-00
Elliot Eisner	98
David Snow	98
Merrill Chandler	98-
Deb Gilman	99-13
Pepe Arostegui	*01-13*
Rattana Buosonte	*01*
Samran Mejang	*01*
Walenia Silva	01-03
Khalil Dirani	02
Izabela Savickiene	02
Juny Montaya	02
Brinda Jegatheesan	03-06
Luisa Rosu	03-11
Ros McPherson.	04
April Munson	05-07
Ivan Jorrín	05-08
Maream Nillipun	09-14
Chotima Nooprick	09
Isabel Arbesu	10
Gloria Contreras	10
Pilar Garcia	11
Nok Wanasathi	11
Shameem Rakha	11-14
Pepe Guererra	12
Fátima Cruz	12-13
Catalina Ulrich	13
Biljana Fredriksen.	13
Sharon Hsiao	13-15
Paulina Cupul	14
Ester Garcia	15
Judit Novak	15-16
Isabel Ramirez	16

Often Remembered

<u>CIRCE Helpers</u>

Clement Adelman
Dan Alpert
Pepe Aröstegui
Mike Atkin
Eva Baker
Kathryn Bloom
Rita Bornstein
Larry Braskamp
Liora Bresler
Bill Brewer
Harry Broudy
Chip Bruce
Al Buccino
Charles Caruson
Eleanor Chelimsky
Cynthia Cole
Nancy Cole
Rita Davis
Gabe Della-Piana
Dick Dershimer
Lizanne DeStefano
Roz Driver
Elliot Eisner
Michael Eraut
Margaret Fiore
Roberta Flexor
Joe Frattaroli
Gene Glass
Egon Guba
David Hamilton
Del Harnisch
Jerry Hausman
Jacquie Hill
Wells Hively II
David Hopkins
Linda Ingison
Dick Jaeger
David Jenkins
Ivan Jorrín
Ken Komoski
David Krathwahl

Saville Kushner
Deborah Laughton
Roger Lennon
Dan Lortie
Bob Louisell
Ulf Lundgren
Barry MacDonald
Martin Maehr
Christine McGuire
Les McLean
Juny Montaya
Jack Morrison
Maream Nilapun
Jeri Nowakowski
Michael Patton
Jim Pearsol
David Pearson
Bill Platt
Jim Popham
George Reese
Fazal Rizvi
Nancy Roucher
Deb Rugg
Michael Scriven
Rudy Serrano
Helen Simons
Lou Smith
Mary Lee Smith
Lawrence Stenhouse
Frances Stevens

Howard Stoker
Dan Stufflebeam
Ken Travers
Ralph Tyler
Decker Walker
Rob Walker
Jim Wardrop
Wayne Welsh
Lyn Wharton
Klaus Witz
Russ Zwoyer