The Impact of CRESST R&D

CRESST Report 523

Ernest R. House, Scott Marion, Linda Rastelli,
Dorothy Aguilera, Tim Weston, and Kyung Min
University of Colorado, Boulder

January 1996

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GESE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

Introduction

This is an evaluation which uses multiple methodologies to estimate the impact of major R&D work in educational testing. Specifically, this is an attempt to gauge the impact of the Center for Research on Evaluation, Standards, and Student Testing (CRESST), headquartered at UCLA. CRESST is a consortium of people and institutions, including partners at the Universities of Colorado, Pittsburgh, Stanford, California-Santa Barbara, Southern California, as well as the National Opinion Research Center and Rand Corporation. The enterprise is funded primarily by the U.S. Department of Education.

For the past five years CRESST's focus has been on conducting research and development on "alternative assessment." Alternative assessment is alternative to traditional achievement testing, long conceived in the United States as primarily pencil and paper, multiple-choice testing, as represented by the Scholastic Aptitude Test, for example. By contrast, alternative assessment consists of such testing modes as portfolios, essays, experiments, performance tasks, oral exams, or pretty much any other way of testing educational performance presumed to be closer to the knowledge or task being assessed.

CRESST has researched, developed, and disseminated knowledge and products about alternative assessment from a wide variety of perspectives. More than eighteen researchers have been involved in some aspect of CRESST work, and others have attempted to implement it throughout the country. CRESST is very much an

umbrella organization in that it encompasses diverse activities and people. This variety of effort makes its evaluation more difficult. What these combined evaluation studies attempt to do is estimate the influence or impact that CRESST has had over the past five years of funding.

There is a tendency to define "impact" as the ultimate effect that one is seeking and hinge the success of the project on the attainment of that ultimate effect. But with R&D, ultimate impact may be very far down the line indeed, and it may take multiple forms. The eventual impact one seeks may be years or decades into the future. Hence, we use the term "impact" more or less synonymously with "influence" or "effects" to indicate shorter or intermediate range impact, as well as effects not anticipated. CRESST influence is conceived broadly over years and populations. Of course, ultimately, CRESST would want to see improved learning on the part of students across the country as a result of alternative assessment.

The first chapter is a rationale for the evaluation, which discusses the particular problems of evaluating R&D work. The second is a bibliometric study of CRESST influence on the research community, the third a national survey of CRESST influence on test directors across the country. The fourth chapter examines the use of two major CRESST products, and the fifth discusses the prospects for adoption by teachers, based on the experience of a pilot project. The sixth chapter draws these findings together into an overall judgment about CRESST impact, summarized briefly in the Executive Summary.

Eva Baker, at UCLA, and Bob Linn, at the University of Colorado, are co-directors of CRESST, and we thank them for their patience and indulgence in this effort to evaluate the impact of CRESST R&D. Anyone being evaluated is a little nervous, and there is more uncertainty in such unfamiliar terrain as R&D evaluation, which is relatively new, except for traditional peer review procedures. We think we have managed to draw a reasonably accurate portrait of CRESST influence, albeit a complex one which has some parts missing.

Contents

Executive Summary

In an attempt to evaluate the impact of the work of CRESST, the Center for Research on Evaluation, Standards, and Student Testing, "effects" were divided into anticipated and unanticipated, with the evaluation focusing on the former. Critical communities of influence were identified, including measurement researchers, test directors, recipients of CRESST products, teachers, and the public. CRESST influence on different groups was evaluated using different research methods for each. "Impact," "influence," and "effects" over a five-year period were conceived more or less synonymously for the purposes of this study.

An extensive bibliometric study using publication and citation indices provided evidence that CRESST researchers produced large numbers of publications, including books, book chapters, technical reports, and articles, many of which were published in the highest status journals. These publications were cited frequently by other researchers in the field and constituted a significant portion of the publications on the topic of alternative assessment. A "use" indicator determined how centrally the cited publications were used in the articles of citation. According to all these analyses, CRESST had a very substantial influence on the measurement research community.

A second study focused on test directors and others who are important in district and state testing programs. This significant gatekeeper group believed that alternative assessment provided an important improvement to traditional assessment, according to the national survey. However, alternative assessment should only be

used in addition to traditional assessment, in their opinion. They thought that CRESST had been significantly helpful in thinking about and decisions to implement alternative assessment. CRESST produced research of high quality and served as a reliable and objective source of information, in their opinion. CRESST influence was exercised mostly through impersonal modes of communication, such as journal articles and technical reports.

A third study traced the distribution and use of two major CRESST products, which were nominated by CRESST. A book on how to do alternative assessments was distributed to tens of thousands through a professional curriculum association and used throughout the country to train teachers and administrators. The book was highly praised for its clarity, utility, and high quality, and for addressing an important need. The use of the CRESST cognitive model was more equivocal. Several groups have begun using the model but comments were too preliminary to make a judgment about the model's effectiveness at this time.

A fourth study focused on teachers. Alternative assessment has not been implemented by teachers across the country to any significant degree so far. However, a pilot project in which CRESST researchers worked with teachers to develop alternative assessments for their classrooms indicated that participating teachers perceived substantial benefits from alternative assessment, especially in improving performance assessment, involving students in their own assessment, and diagnosing student learning problems. However, developing such measures required a huge amount of time and effort by both teachers and developers. And

the new ideas did not spill over to other faculty in the schools during the three year period. Projected across the country as a whole, the implementation of alternative assessment would require an enormous investment of time and resources.

Finally, a review of newspaper articles on alternative assessment from the Lexis/Nexis national data base indicates that there is fast rising public interest in the topic. In 1990 there were only five articles, but by 1990 there were 85, half in local newspapers, most articles discussing assessment policies and controversies. The number of articles has doubled every year since 1990. CRESST was mentioned by name in only a few but since its work is with intermediaries, this probably underestimates influence on the public.

Two major limitations of this evaluation are that no study of CRESST influence on policy makers was conducted, which almost certainly results in an underestimate of CRESST influence since several CRESST researchers have been involved with policies at the national, state, and local levels, and no in-depth case study of alternative assessment functioning fully in a district or state was undertaken, which probably results in an underestimate of the problems with alternative assessment implementation.

In summary, CRESST has made a powerful impression on both measurement researchers and test directors, groups one might anticipate as being critical to the acceptance of alternative assessment, given their professional authority and gatekeeper functions. One must assign CRESST very high marks here. CRESST products have also had a very favorable reception, though one more

qualified than the influence on researchers and test directors. Pilot work with teachers suggests that alternative assessment can be implemented with perceived benefits, but that the cost will be very high across large number of teachers and classrooms. Hence, in the past five years CRESST has had a powerful impact on gatekeeper groups and has indicated what the pathways and costs may be for national classroom implementation.

Evaluating R&D Impact

Ernest R. House, Scott Marion, Linda Rastelli, Dorothy Aguilera, and Tim Weston

University of Colorado at Boulder

Feb., 1996

Few things seem more difficult than evaluating research and development impact. In fact, when the problem is posed to most researchers, their reaction is that it can't be done, except perhaps by peer review. Although peer review should play an important role, it hardly seems sufficient for evaluating R&D "impact" (Chubin, 1994). R&D expenditures average 2.3% of GDP in developed countries and are currently $68.1 billion in the US (Wargo, 1994.) There are building political pressures for evaluating R&D efforts.

Evaluation methods have been unequal to the task. "The uncertainties are too great, the causal paths too diffuse, the benefits too difficult to measure, the time scale too extended" (Roessner, 1993, p. 197). An Office of Technology Assessment study concluded, "Since 1985, no breakthrough methods of any variety have been invented that more definitively reveal the ex post scientific or social value of past research investments...." (H. Averch, 1990, quoted in Kostoff, 1993, p. 175). Most such evaluation is of hard science projects. The evaluation of social science based R&D may present even more difficult problems.

We will attempt to outline a rationale, methodology, and some results for R&D impact evaluation, specifically for the Center for Research on Evaluation, Standards, Student Testing (CRESST). CRESST is the major R&D center developing "alternative assessments." It is funded by the Department of Education and headquartered at UCLA, with

cooperating researchers at the Universities of Colorado, Pittsburgh, Stanford, and other related institutions.

Dimensions of Evaluation

Three dimensions for evaluating R&D are quality, social worth, and ethics. One can have good research which is meritorious on its own terms but which has little or no social impact, and one can have bad research which has quite a lot of impact. Ideally, one would want high quality research with high impact. There is some agreement on procedures for judging research quality, the method of choice being peer review, according to agreed criteria, e. g. consistency with previous research, conclusions consistent with data, and appropriate data collection techniques (Chubin, 1994).

CRESST has been subjected to many peer reviews. If one assumes that the R&D is high quality, what then? So far, the ethical dimension of R&D has been dealt with mostly through university review boards (Howe and Dougherty, 1993; Sieber, 1982). Although the ethical dimension is important, we limit ourselves to impact evaluation, since that seems relatively undeveloped. Shadish (1989) contends that concepts from program evaluation are useful for evaluating R&D. He distinguishes between internal-external and process-outcome evaluation criteria, noting that almost all methods for evaluating science use criteria internal to the discipline. However, program evaluation, from which the unified field of evaluation has emerged, is not the same as R&D impact evaluation.

The special problem is that the effects of R&D projects are so uncertain. Effects may take quite a long time to emerge and the pathways of R&D influence may be indirect, delayed, or obscure. "The uncertainties associated with R&D, its multiple consequences, cumulative nature, and

transferability all help to explain why the evaluation of R&D is so difficult....A critical problem in evaluating R&D activities is the long and uncertain time frame in which 'results' may be observed" (Melkers, 1993, p. 44). This pervasive uncertainty about "payoff" far down the line makes R&D impact evaluation difficult.

Anticipated vs. Unanticipated Impact

We are using the terms "effects," "impacts," and "influences" interchangeably, though they carry different connotations. Impact is the favored term these days, perhaps because it reflects direct, simple, unidirectional, almost physical, effects. "Influence" suggests something more tentative, perhaps two-way effects mediated and modified through groups of people. "Effects" is the most general term and suggests that there may be far-reaching, unanticipated, diffuse, and undetected things happening.

One way to approach this problem is to separate potential R&D effects into the unanticipated and the anticipated. We might admit that ultimate long-term effects of R&D are beyond our present ability to assess, though one might be able to determine unanticipated effects by retrospective case studies. For example, R&D "spin-offs" might involve applications of ideas, products, or technologies not planned or anticipated at the beginning (Brown and Wilson, 1993). These may occur by application of the idea to new markets or domains not anticipated, the classic case being military R&D adapted to civilian uses.

New technologies may be reworked or combined with other ideas to produce "second-generation" technologies. In other words, new technologies differ greatly in their "robustness" or ability to generate further (unanticipated) ideas down the line. These second-generation

technologies often grow from the failures and learning experiences of the original project. Sometimes they are linked to "enabling" technologies that solve critical technological or marketing problems. Although most spin-offs are accidental by-products, they significantly enhance the payoff from the original technology (Brown and Wilson, 1993).

On the other hand, there are R&D effects, impacts, or influences that are foreseeable. That is, for particular types of R&D one might expect to find effects more likely with some groups rather than others. In basic research one might expect to find responses from colleagues at the theoretical level, but not expect to find marketable products. If this is so, tracing influence by bibliometric techniques for specially-defined populations would make sense for basic research, but not for applied research.

Weiss (1989) has said that one should distinguish between "knowledge utilization" and "innovation diffusion." Most research does not produce innovations and most innovations are not derived from research. She suggested that one should define potential clients broadly, including policy makers and the public, that one should think in terms of intermediate organizations by which knowledge can be diffused, and that one should think less in terms of discrete studies than in terms of compilations of evidence on particular subjects.

One R&D impact evaluation of hard science technology, unusual for its scope, consisted of 31 retrospective case studies of technology absorption and transfer for the New York State Energy Research and Development Authority (Kingsley, Bozeman, Coker, 1995). This retrospective "aggregate case approach" compared data across cases qualitatively and quantitatively. Two change processes emerged from

these cases, those in which "technology absorption" into cooperating agencies occurred and those of "technology transfer," in which third parties removed from the development used the results for their own purposes.

Where no one used the technology, it was not necessarily for lack of trying on the part of developers. Ordinarily, no use occurred because a key actor withdrew support for a number of reasons, including belief that the market wasn't there, that the technology wasn't good enough, or because of local politics. Many projects involved risky prototypes and were sponsored by sole source funding rather than competitions. Hence, evaluating effects solely in terms of transfer efforts can be misleading since much effort may lead to no effects. Bozeman, Papadakis, and Coker (1995) also studied relationships between federal R&D (hard science) labs and commercial corporations. These federal labs were free-standing, government-sponsored enterprises working directly with companies to develop products.

Methods for Data Collection

In the literature on evaluating R&D impact, here are the major data collection methods (omitting patent analysis):

Return on investment--Used in business and economic analyses sometimes. One must calculate benefits and costs (Link, 1993). There are cases where precise estimates are not needed, e. g., a project is costly but has little or no payoff, but these methods have not been applied with much success, even though this is the kind of information legislators would like to see. (Bozeman, Papadakis, and Coker, 1995, p. 44, described the R&D evaluation environment as one of "desperately seeking numbers.")

Although it is easy to imagine such evaluation, e. g., put an R&D product into place and measure the increase in achievement test scores related to costs, these measurements are beyond our capabilities currently. In fact, this type of analysis was the intent of the Follow Through program, with all its attendant difficulties. The dozen or so Follow Through early childhood education "models" produced so many different results that there was as much variance of student achievement within each model as between models. In other words, the same model at different sites produced quite different results. Such contextually sensitive results does not provide a stable basis for assigning costs and benefits across sites.

In fact, defining production functions for education has not been successful generally (Monk, 1992). There are several reasons for this, the most likely being that educational production (achievement by students) is caused by an array of interactive factors so that introducing a new educational technology leads to different results, depending on what the factors are in the context. One can evaluate success within context but generalizing across contexts is more difficult.

This disappointing result has not been limited to educational programs. In a review of US federal R&D evaluation (hard science and technology), Kostoff concluded, "Cost benefit analysis has limited accuracy when applied to basic research because of the quality of both the cost and benefit data due to the large uncertainties characteristic of the research process, as well as selection of a credible origin of time for the computations" (Kostoff, 1993, p. 174). R&D evaluation has relied heavily on peer review procedures instead.

Not even business firms heavily engaged in R&D use quantitative measures to evaluate their productivity. In one study only 20% of the leading 34 R&D firms used any quantitative productivity measures; 59% used none. One R&D firm director said, "...attempts to quantify benefits of R&D have led to monstrosities that caused more harm than good" (Roessner, 1993, p. 188). This approach doesn't look promising at the moment.

Bibliometric methods, such as ascertaining the number of publications and citations for authors or groups of authors, might be useful for judging the success of basic research. Number of publications is an indicator of productivity or output more than quality, while citations may reflect quality to some degree. However, bibliometric experts caution against using citations to judge individual scholars because the margin of error is so great. Citations are better applied comparatively to similar groups. Citations may also be more indicators of use than quality (e. g., Toffler's Future Shock was the most frequently cited scientific publication over one time period but is unlikely to influence scientific research very much).

Shadish (1989) notes that publication and citation indices are the only widely accepted indicators of scientific quality. He found that highly cited works are judged to be of high quality but that most high quality works are not cited frequently. Awards, honors, and research grants received can add to a mixed list of indicators, but one must be cautious about adhering strictly to such indicators (Shadish, 1989).

Surveys are useful for tracing influence in large populations, assuming that the type of influence is not too complex. For example, Stalford and Stern (1990) carried out a survey of school districts in the

US, asking whether the district superintendents had heard of products from the US Department of Education regional labs, R&D centers (such as CRESST), and the ERIC system, and if they had, what they had done with the information. The survey covered a sample of potential users of R&D products. The limitation was that "influence" was expressed in simple terms on one page.

In a study of government science and technology evaluation in the Canadian government by the Office of Comptroller General, 90 percent of the evaluations used client or stakeholder surveys, 40 percent used surveys of expert opinion, 67 percent used literature reviews, 26 percent case studies, and 13 percent non-client interviews. According to the study, successful client surveys included experts in the survey process, segmented survey respondents, and used experienced interviewers well-versed in the subject to probe the issues. Surveys are useful, but also insufficient.

Case studies are better at tracing complex events, including the most diffuse and complex outcomes, but they have problems of subjectivity, sampling, and comparability (Stake and Easley, 1978; Kingsley, 1993; Stake, 1995). Also, they are expensive to conduct so that their number must be limited. However, there probably is no better way to assess complex influences over a period of time. Of course, even if one employs case studies, the problem of how long it takes for R&D to register effects remains a problem.

Where case studies have been used in R&D evaluation, they have been effective in generating specific information and giving managers a "feel" for the programs. In general, "soft" techniques have been deemed more appropriate because of the abstract nature of R&D programs (Barbarie,

1993). One might note that, "In business as in government, evaluation of basic research is recognized as necessarily a judgmental process, best accomplished through use of informal, largely qualitative methods" (Roessner, 1993, p.199). The New York State Energy 31 case studies offer an intriguing methodology.

Network analysis is another possibility, though not one that has been tried. If one conceptualizes influence as being exercised through personal networks of people and contacts, one can apply methods for analyzing the networks themselves, including intermediate pathways through which R&D ideas and products might exercise influence, and hence anticipate success as a function of the networks in advance of utilization, thus circumventing the time lag to some degree.

For example, there is extensive evidence in the innovation diffusion literature that face-to-face interaction is critical to the diffusion and acceptance of innovations. Similar findings have emerged in the technology diffusion literature, with one study claiming that national research performance stems not only from the activities of individuals but also from their interactions, with some countries having distinct advantages in this regard (Dalpe and Anderson, 1993).

Of course, R&D networks can consist of many things: connections within R&D facilities; connections among potential adopters or users; connections among R&D organizations; connections among members of user organizations; connections between journals and their readerships; invisible colleges; newspaper readership and community influence organizations, and so. Hypothetically, one could define networks of influence for particular projects. The more R&D ideas are diffused to the general public, the more these networks may function like impersonal

mass communication. The more the technology requires extensive changes in behavior, the more the diffusion may require face-to-face interaction.

The advantage of conceptualizing R&D influence as exercised through personal networks is that the process could be subjected to precise analyses, even mathematical analyses, that can be applied to networks generally, whether these networks are based on face-to-face contacts or attitude change through mass communication (Hägerstrand, 1967; House and Long, 1974; Hood, 1989; Mill and Stephens, 1992; Zaller, 1992). Using the concept of the network turns the unknown into something more familiar and analyzable.

In summary, our ideas for evaluating R&D impact are to divide the effects into anticipated and unanticipated and to focus on the anticipated. One might anticipate R&D work will have impact on certain populations and that some influence will be carried through networks of people engaged in face-to-face interactions. Other influences might be carried through impersonal networks, such as the media. Hence, one could define network populations and trace influence, and possibly define characteristics of networks that would anticipate and facilitate success in advance of ultimate effects. For example, one might predict that a particular network would be unlikely to diffuse an R&D idea because of its structure.

The Case of CRESST

CRESST is not a single-dimensioned entity, but a consortiuum of well-known scholars who cooperate voluntarily. Their activities are many and diffuse. Such a diverse structure poses a problem for evaluators, however, in that effects may be generated from many sources to many different groups. The CRESST program for researching and developing

"alternative" modes of student assessment consists of work along these lines (Dietel/Herman memo, Feb. 2, 1995):

A. Assessing assessment: assuring the quality and validity of assessment systems (e. g., the CRESST criteria);

B. Effective models for developing alternative assessments;

C. Systems and implementation strategies that serve accountability and improvement;

D. Approaches to assuring technical quality;

E. Technical methodologies to optimize cost effectiveness;

F. Models for standards setting;

G. Models for assessing and implementing opportunity to learn; assuring equity in assessment;

H. Effective reporting formats;

I. Effective professional development models;

J. Costs and effects of large-scale assessment;

K. Lessons in state assessment policy.

The purpose of CRESST is to develop new and better ways of student testing, especially focused on "alternative assessment." One might expect the effects of such R&D work to be registered closer to educational practice than theoretical work might be, though some CRESST research is theoretical. Again, one might divide influence into the anticipated and unanticipated. Anticipated effects would include influence on groups and individuals with whom the Center is actively working and targeting, including,

--those who attend conferences sponsored by CRESST;

--those who order products or other information from CRESST;

--those who work directly with researchers and developers;

--readers of journals in which the ideas are published;

--other R&D specialists in the same fields;

--targeted "user" populations, such as teachers, school districts, state education agencies, and test developers;

--the public at large, or some segment of it.

These groups can be ordered as to the type of influence one might expect. One would expect policy makers, test makers, and the public to react in different ways to CRESST R&D than do researchers. Some groups operate closer to the R&D work and with the original concepts, products, and researchers, while other groups are far removed and only distant potential users, such as the vast majority of teachers, administrators, and the public. Nonetheless, the teachers and students are presumably the ones ultimately affected by these efforts.

Another way of characterizing these relationships is that researchers and developers exert their influence through established networks, such as those defined by journals or conferences, or by trying to establish new networks, such as by building enduring contacts with those who want to work with the ideas. Hence, one might characterize influence activities partly as network building. The advantage of conceptualizing the problem this way is that one can investigate anticipated pathways of influence by thinking of network construction, thus introducing an intermedicate set of constructs that can be analyzed for potential influence, i.e, the characteristics of the networks themsevles.

Another complication is the unit of influence. Should it be the persons doing the R&D, the project or program, or the ideas, concepts, and materials that are the focus of study? The answer would seem to be, "It depends." In the early days of R&D it would make sense to trace the

influence of particular persons. The originators' personal influence is likely to be registered more at the beginning. If there are several people working on the same ideas, the program or project would seem to be an appropriate unit of analysis. However, sometimes it is not clear where one project stops and another begins.

As ideas and products spread, however, contact with the originators or even the projects may become less likely so that the ideas, concepts, and products, e. g., "alternative assessment" in a vague sense, may be the way change is registered, without reference to sources. By a later time influence may be spread through so many intermediaries that its sources may not be known. The ideas may also be transformed and renamed, and/or attributed to incorrect origins. In fact, the manner in which educational research findings are used by policy makers, use based on third or fourth hand information or media accounts, is similar to the formation of mass opinions (Zaller, 1992).

For example, perhaps the most successful concept in educational R&D over the past few decades is meta-analysis (Glass, Smith, McGaw, 1981), the use of which has spread far beyond its original uses in education. Medical research in particular has made extensive use of meta-analysis, but it is often the case that the technique is claimed to have been invented by researchers in medicine. Such obfuscation is common.

Applying Concepts and Methods to CRESST

How then have we evaluated the impact of CRESST R&D? We divided the potential influence/impact groups into the following:

Researchers--To assess the impact on researchers, we conducted analyses of publications, including numbers of publications of major CRESST researchers, numbers of citations, most important publications

(in the judgment of CRESST), how these citations were used, and what kinds of journals these citations appeared in. We used the Social Science Citation Index as our main data source, but also counted publications in important educational journals not included in the citation index, such as the Educational Researcher. Including use and journal status moves beyond regular citation analysis.

Test directors--We surveyed members of the national association of test directors through the mail, asking them what influence the CRESST products and ideas had on their attitudes and behavior. We conceived test directors to be an important audience because of their gatekeeper function regarding testing, even though many had little contact with CRESST.

Practitioners--Practitioners consist of . It is not likely many administrators and teachers are well informed about alternative assessment at this time. The ideas are too new. In the Stalford and Stern (1990) study, recognition of federal R&D institutions was high (over 90 percent). Of the 64% of the districts recognizing federal R&D centers, 52% received products, with the majority of these (37%) reporting frequent use. However, when asked about "particularly useful" materials the R&D centers received only 2% mention. In general, the Department of Education evaluators were surprised by the positive responses. In another study the federal R&D centers and laboratories were judged not effective in disseminating their work (Center for Leadership Development, 1984).

We conducted two studies to trace the influence of two CRESST products. Telephone interviews were conducted with people nominated by the developers, then with people nominated by the interviewees, and so on, until a chain of influence was traced from developers to use.

Teachers--Usually, teachers are considered the ultimate audience for CRESST impact, yet few projects have worked directly with teachers. One might expect that it would take quite a long time for a substantial number of teachers to use CRESST products. CRESST research reports were analyzed to gain an idea of how teachers are likely to react to alternative assessment. Interviews were held with a few teachers who developed alternative assessments for their classrooms.

Policy Makers--Within this evaluation, there are some important things omitted. First is the impact on policy makers. Certain CRESST people worked with policy makers to set testing policies at the state and national levels. However, when we solicited from CRESST researchers what influences they thought they had, the nominations were erratic. Most did not name policy makers Hence, we omitted this influence because we could not see how to bound it systematically. It also seemed unlikely we could survey policy makers with mailed questionnaires or telephone surveys, the limits of our resources.

Case Study--We also considered overall case studies. One can have a number of pieces but these may fit together in unexpected ways because politicians, tests, teachers, etc., interact with each other in powerful ways. The ultimate challenge is to see how everything works together in a school district or state. We were never able to settle on a state or city that represented full alternative assessment implementation.

The public--One might expect CRESST to inform the public as well as professionals. The Nexes data base was searched to ascertain how many times alternative assessment appeared as a topic in newspaper articles.

Network Studies. We did not conduct extensive studies along the networking lines suggested earlier, mostly because of lack of resources. One might define potential networks on which the new ideas and products might travel and analyze these networks, enabling an estimate of impact before the influence had occurred. Analyzing impact on the research and test director communities (networks) was one step in this direction. But one could analyze the journal and director networks in detail.

In summary, one might imagine an R&D technology ("alternative assessment") sweeping the country from its center of creation to transform education. However, that is not how R&D impact occurs. Rather, the R&D center works with diverse groups simultaneously (and opportunistically) to register different kinds of influence within each. If development and diffusion of the technology across the nation is the idealized picture, this evaluation of impact is limited to penciling in discrete segments of the picture so that one has an idea of shape and extent.

Some Results

In this evaluation of impact, we have assessed the influence of the CRESST R&D work on several populations over a five-year period. Although the ultimate aim of CRESST is to improve the testing of students in such a way as to enhance learning, we would not expect CRESST efforts to have resulted in widespread application in classrooms or increased student achievement throughout the country after only five years.

To judge influence on the educational measurement research community, we conducted analyses of CRESST publications and citations, including how many publications were produced, the status of the journals in which they were published, how often these publications were cited and

in which journals, and how citations were used in the context of the article of citation. According to all these indicators, CRESST researchers had a very substantial impact on the educational measurement community. They produced a large number of publications in the highest status research journals and had their work cited frequently by other researchers in ways central to the development of the ideas in the articles of citation. CRESST also published articles in practitioner journals as well.

For example, the core group of CRESST researchers produced 90 articles, books, book chapters, and technical reports that were cited 424 times between 1990 and 1995, or 4.7 times per cited article. This is a substantial number by almost any standard. Furthermore, the articles were published in many of the highest status journals in the field, indicating acceptance by peer review and access to the leading scholars. Although self-citations (17%) and CRESST partner citations (24%) were siginificant, the majority of citations (59%) were by researchers <u>not</u> connected with CRESST. Of course, these citations were not normally distributed. A few publications garnered most of the citations.

On a three-point rating system of journal status, the CRESST publications rated 2.3 on average. This high status ranking was achieved in spite of the fact that many CRESST researchers published articles in lower status practitioner journals in order to influence practice, thus bringing down their overall scores, a limitation of our indicator.

Use of the articles cited was also analyzed. For example, the Linn, Baker, Dunbar (1991) article had the most citations among key CRESST publications, with an average citation rate of 13.1 per year. On a three point use scale from least to most important use, it's average was 1.8. In general, this number reflected how most CRESST research was used in

citations. A use rating of "3" indicates that the work was "critical" to the article in which it was cited, and "1" indicates that another publication could have been used in its place, even though it was cited. So centrality of use was substantial for the nominated CRESST articles analyzed.

To determine the boundary of CRESST influence on the measurement community, an ERIC-CJE search for 1994 and 1995 discovered 54 performance assessment articles, 10 of which were published by CRESST partners, or 18.5% of the total, a substantial portion to emanate from one research program. In an examination of 35 articles we could find, CRESST research was cited 90 times in 9 CRESST authored publications and 42 times in 26 non-CRESST articles. Most CRESST articles were in the highest status journals. All in all, the evidence is extensive that CRESST has had a major impact on the measurement research community.

One limitation to this bibliometric study is that there are no comparable groups of researchers to compare to those in CRESST, other than those in fields in which the publication and citation practices are different. However, even without such comparisons, the volume of publications, citations, and uses is so high that influence and impact on this particular research community are unmistakable.

The second community of influence was that of test directors, publishers, and others who serve as gatekeepers to district and state assessment procedures. A national survey of this population revealed the strong influence of CRESST R&D on them as well. The test directors were convinced that alternative assessment was important, that it was a significant improvement in assessment procedures, and that CRESST was a major, credible, and highly valued source of information on the topic. Not only did the directors agree that CRESST was a major influence on

their thinking and their decisions, but they lauded CRESST for the high quality and objectivity of its work.

As a group, test directors were open to alternative assessment techniques, with writing assessments, performance tasks, and portfolios being the most popular, and exhibitions, experiments, and oral exams the least popular. These alternative forms of testing were used at a high rate, albeit at a rate less than that of traditional, standardized tests. Of course, a few directors did not like alternative assessment at all, mostly because of its perceived lack of validity and reliability.

There are important qualifications. First, the test directors accepted alternative assessment techniques only as a supplement, not a replacement, for traditional standardized achievement testing. For the most part they saw alternative assessment as useful at the classroom level to improve teaching. It was not seen as useful for accountability at the district or state level. Traditional achievement testing was perceived as better for that. The positive attitude of test directors towards alternative assessment probably would change if they were forced to choose between alternative and traditional assessments. One must wonder how teachers having to teach for both traditional and alternative assessment will affect their classroom behavior. Clearly, the demands on them will be much greater.

Test directors saw CRESST as an important and influential source of information on alternative assessment. About 30% of respondents saw CRESST as very useful, 31% as useful, and 24% as somewhat useful. Only 4% saw CRESST as not useful at all as a source of information. In open-ended comments directors were extremely laudatory about the quality and objectivity of CRESST information and research, perceiving CRESST as a valued and reliable source.

Most directors' previous experience with alternative assessment was with writing assessments, though alternative assessment is a recent experience (the last two years) for half. The directors relied on many other sources of information other than CRESST, so their attitude cannot be attributed solely to CRESST. Impersonal sources of information, such as journal articles, seemed to dominate their contacts with CRESST, though about one-quarter had personal contacts. One can conclude from this survey that CRESST has had a significant and continuing impact on this important gatekeeper community.

A third study examined the influence of two CRESST products to see both how and how widely the products were used. Products are much closer to the practitioner community than journal articles. CRESST was asked to nominate two of their best products, and we tracked how these products were used in the field by phone interviews. Admittedly, these tracer studies have a positive bias because we asked CRESST to nominate two of their best products and suggest the names of those professionals who had made use of these products.

The first product was the book, A Practical Guide to Alternative Assessment by Herman, Aschbacher, and Winter. This publication was distributed to 90,000 members of the Association of Supervision and Curriculum Development (ASCD) through their regular publication list. ASCD is an organization of curriculum directors and others in school districts who are responsible administratively for curriculum matters at the district level. Clearly, this is a huge distribution. CRESST distributed another 44,000 copies through other sources.

Telephone interviews with those who have used the book revealed a high regard for the product's quality and usefulness. Mostly, the book was

used for training teachers and administrators in alternative assessment techniques, and such a product was badly needed in the field, respondents reported. The state of Illinois used the book extensively as part of school planning processes. Most users thought the book covered the essential topics in easy to understand language and was put together so individual chapters could be used.

Using the existing ASCD network helped the distribution considerably. Most said they had been looking for something on alternative assessment to use before discovering the book. Often, product use started with a state department of education initiative, which stimulated districts into action. Word of mouth was a favorite dissemination pathway, with most users saying they had recommended the book to 25 to 100 people. Brevity and simplicity were perceived positive attributes.

The other product, a content assessment model based on cognitive psychology, has been used by Hawaii and Los Angeles schools. However, these and other projects in Missouri and Washington are not far enough along to evaluate definitively. Users have tried to develop assessments based on the model, but, in general, the model requires complex implementation and adaptation. In the case of both the book and the model, users were already aware of alternative assessment but needed tools for implementing it.

The final study was an attempt to assess impact on teachers, the ultimate group who must implement alternative assessments. At this time, relatively few teachers across the country have tried these techniques. The techniques are too new. Hence, there was no sense in conducting a national survey. Yet teachers are the most critical group of

all (except students), and the success of alternative assessment must rest with their eventual acceptance and use.

To address these problems we examined one of CRESST's pilot projects. CRESST has undertaken a few pilot studies in which a small number of teachers were helped to develop alternative assessment measures for their classrooms. We examined the documents produced by these pilot projects to estimate the benefits and problems that teachers across the country are likely to encounter as they implement these new measures. We interviewed a few participating teachers independently after the pilot project was over.

In general, participating teachers adopted the techniques that fit their underlying beliefs, but not those that did not. Pre-existing beliefs of individual teachers towards instruction and assessment and their reliance on text books turned out to be significant factors in implementation. The development also took a great deal of time and effort, and some teachers thought that the effort detracted from time spent on instruction. This necessitated a reduction in workload eventually. "Comp time" amounting to a full day a month would have helped ease the extra burden, in their opinion. Also, other projects on-going at the schools, unrelated to CRESST, made implementation more difficult.

A second year of project participation was needed, and even then only by the third year did teachers feel comfortable with the new ideas. Nonetheless, the participating teachers thought they had gained by improving the performance assessment of students, involving students in their own assessment, and diagnosing problems students faced, which included being able to discuss student performance better with parents.

In general, assessement was more integrated with classroom instruction. Participating teachers varied individually in what they did and how they did it, including their acceptance of the new ideas. Apparently, the new ideas did not spread to other faculty in the schools over the three year period.

Finally, according to information from the Lexis/Nexis news service, since 1990 there has been a rapidly increasing number of articles on alternative assessment. In 1990 there were five articles, two in the New York Times, two in other major newspapers, and one in an educational publication. In 1994 there were 85 articles, 42 in regional or local papers, 22 in educational publications, 18 in mass circulation magazines, and 11 covering Congressional testimony. First, major newspapers recorded the trend, picked up later by local and regional papers, followed by Congressional and legislative sources. Most articles described new assessment policies and the controversy over their introduction.

There are at least two significant limitations to this evaluation. First, we did not ascertain CRESST influence on policy makers. In our view, not including policy makers underestimates CRESST influence because several CRESST researchers have worked with policy makers at the district, state, and national levels. The second limitation is that we did not conduct an in-depth case study of a large unit to see how alternative assessment interacts with other factors. Innovations always change in the course of their implementation. We probably have underestimated the problems that will arise in the implementation.

Finally, what can one conclude about CRESST influence, impact, and effect over this five year period? First, CRESST has had a very powerful influence on two significant reference groups, researchers and test

directors. Although these are not the groups who must do the alternative assessments, it is difficult to imagine progress without their acceptance, support, and participation. CRESST influence was achieved mostly through publications and impersonal contacts. One must assign CRESST very high marks on impact here. Again, we note that this is alternative assessment seen as supplement, not replacement. It is also true that CRESST researchers probably find influence exercised through publications most compatible with their academic style.

CRESST products have also made a significant contribution. One product has been very widely disseminated and used, especially in teacher training. CRESST relied heavily on a pre-established network, thus leveraging influence and minimizing costs. For the second product, it is not clear at this point whether the CRESST model has been successfully implemented. On products, we would give a good but qualified grade overall.

Finally, for teachers the experience has been positive, but with sobering qualifications. Very talented researchers and developers, including four principal investigators and four graduate students, worked with fourteen teachers to develop and implement alternative assessments in one pilot project. This implementation required a substantial investment of personal time and contact. The development was not easy or quick. For the most part the teachers were satified afterward about the use of alternative assessment in their classrooms, even while admitting that the effort invested was very substantial. The new ideas did not spread to other faculty.

In general, problems become more numerous the closer one moves to implementing alternative assessments in classrooms. If one looks

across the country and imagines the time and effort required to implement alternative assessment in hundreds of thousands of classrooms, then the investment will have to be huge. One must wonder where these resources will come from and how many decades such a change might take.

Conclusions

Evauating R&D impact is not easy, but it is not impossible. The approach we have developed here is to define communities of importance to the ultimate impact of the R&D and to estimate the effects on those populations, keeping in mind the ultimate goals. This circumvents the difficulties of asking for impossible results immediately or of assessing the immediate effects far removed from the ultimate goal. To evaluate the impact on different populations probably requires different data collection methodolgies, since the effects of different activities are registered differently.

In the future one could imagine not only multiple methodologies but also techniques for investigating the characterisitics of networks that facilitate or impede impact within particular communities. For the most complex effects, however, one probably still must resort to retrospective case studies.

References

Averch, H. A. (1993). Criteria for evaluating research projects and portfolios. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 263-277.

Barbarie, A. J. (1993). Evaluating federal R&D in Canada. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 155-162.

Bozeman, B. (1993). Peer review and evaluation of R&D impacts. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 79-98.

Bozeman, B., Papadakis, M., Coker, K. (1995). Industry perspectives on commercial interactions with federal laboratories. Atlanta, GA: School of Public Policy, Georgian Institute of Technology. 75 pages.

Brown, M. A. and Wilson, C. R. (1993). The temporal dimension of R&D evaluation: Incorporating spin-off benefits. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 244-262.

Center for Leadership Development. (1984). Creating and disseminating knowledge for educational reform: Policy management of Education's regional educational laboratories and national research and development centers. A report to the National Council on Educational Research. Los Angeles, CA. 116 pages.

Chubin, D. E. (1994). Grants peer review in theory and practice. Evaluation Review, Vol. 18, 1, 20-30.

Cozzens, S. E. (1993). US. evaluation of strategic research: Closing the gap with Europe and Asia. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 229-244.

Dalpe, R. and Anderson, F. (1993). Evaluating the industrial relevance of public R&D laboratories. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. (207-228).

Hägerstrand, T. (1967). Innovation diffusion as a spatial process. Chicago, IL: University of Chicago Press.

Hood, P. D. (1989). How can studies of information consumers be used to improve the education communication system? Paper presented to the American Educational Research Association, March 29, San Francisco, CA.

House, E. R. and Long, J. M. (1974). Applying directed graph theory to faculty contact structure. Paper presented at American Educational Research Association, Chicago, IL, April 19.

Howe, K. R. and Dougherty, K C. (1993). Ethics, institutional review boards, and the changing face of educational research. Educational Researcher, 22, 9, 16-21.

Kingsley, G. (1993). The use of case studies in R&D impact evaluation. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers, 17-42.

Kingsley, G., Bozeman, B., Coker, K. (1994). Technology transfer and technology absorption: An aggregate case approach to evaluating RD&D impacts. Syracuse, NY: Center for Technology and Information Policy. Syracuse University. 61 pages.

Kostoff, R. (1993). Evaluating federal R&D in the United States. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 163-178.

Link, A. N. (1993). Methods of evaluating the return on r & d investments. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 1-16.

Melkers, J. (1993). Bibliometrics as a tool for analysis of R&D impacts. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers, 43-61.

Mills, Stephan R. and Stephens, K. Gwen. (1992). Impact of education dissemination practices on California schools. Far West Lab, San Francisco, CA. ERIC Document 354878.

Monk, D. H. (1992). Education productivity research: An update and assessment of its role in education finance reform. Educational Evaluation and Policy Analysis, 14: 4, 307-332.

Roessner, D. (1993). Use of quantitative methods to support research decisions in business and government. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 179-205.

Shadish, W. (1989). Science evaluation: a glossary of possible contents. Social Epistemology, 3, 189-202.

Shadish, W. R. (1989). The perception and evaluation of quality in science. In B. Ghoulson, W. R. Shadish, Jr. R. A. Niemeyer, A. C. Houts. (Eds.). Psychology of science: Contributions to metascience. New York: Cambridge University Press. 383-426.

Sieber, J. E. (Ed.). (1982). The ethics of social research. 2 vol. New York: Springer-Verlag.

Stake, R. E. and Easley, J. A., Jr. (Eds.). (1978). Case studies in science education. Center for Instructional Research and Curriculum Evaluation. Urbana, IL: University of Illinois.

Stake, R. E. (1995). The art of case study research. Thousand Oaks, CA: Sage Publications.

Stalford, C., and Stern, J. (1990). Major results of a survey on the use of educational R&D resources by school districts. American Educational Research Association, Boston, MA, April 16-20. ERIC Document 329 212.

Wargo, M. J. (1994). A congressional agency's evolving evaluation approach to research and development. Paper presented to American Evaluation Annual Meeting, Boston, Nov. 4, 1994.

Weiss, C. H. (1989). Improving the dissemination of research, statistical data, and evaluation results in education. Cambridge, MA: Harvard University. Mimeo.

Zaller, J. R. (1992). The nature and origins of mass opinion. Cambridge, UK: Cambridge University Press.

Chapter 1

Evaluating Research and Development Impact: The Case of CRESST

Ernest R. House

Few things seem more difficult than evaluating research and development efforts. In fact, when the problem is posed to most researchers, their first reaction is that it can't be done, except perhaps by a peer review panel which estimates potential benefits of R&D efforts in a process similar to reviewing research proposals. Although peer review certainly should have an important role in evaluating R&D, it hardly seems sufficient for evaluating the "impact" of R&D. Researchers who pride themselves on empirical testing cannot be satisfied with an inability to discover the effects of R&D, difficult though such a task might be.

Political, structural, and methodological problems have impeded such evaluation in the United States. From the Second World War through the Cold War, science was acclaimed and funded as critical to the national defense. The great technological successes of WW II convinced both the government and public that scientific research and development were extremely important to the national interest. Science and engineering departments in major universities were built on expenditures for defense-related research (Leslie, 1993). In a sense, the fruitfulness of R&D was taken for granted.

However, times are changing. There are strong political pressures for evaluating R&D efforts (Wargo, 1994). Although science is still highly regarded, it is no longer sacrosanct, and a

less friendly eye has been cast on R&D expenditures, especially during a time of decreasing federal budgets, the source of most R&D funding. Although R&D is still considered important to the national economy, it is being subjected to closer scrutiny and cost calculations. What is its payoff? What is its worth?

There are also structural reasons why evaluation of R&D has been undertaken with less vigor in the US than in Japan and Europe. "Government-owned" research institutions are more prevalent in those countries, whereas US research activities are highly decentralized, e. g., in universities, which have their own evaluation systems not necessarily focused on social or economic worth (Cozzens, 1993). US research has been guided by grants to individual institutions, rather than by large block grants (with some exceptions). Centralized control in other countries has led to more formal evaluation, often to improve the management and efficiency of such expenditures. R&D expenditures now average 2.3% of GDP in developed countries and are currently $68.1 billion in the US (Wargo, 1994.)

Finally, R&D endeavors have not been evaluated because evaluation methods have been perceived as being unequal to the task. "The uncertainties are too great, the causal paths too diffuse, the benefits too difficult to measure, the time scale too extended" (Roessner, 1993, p. 197). Researchers themselves feel this way. A 1990 Office of Technology Assessment study concluded,

> Since 1985, no breakthrough methods of any variety have been invented that more definitively reveal the ex post scientific or

> social value of past research investments....The evidence is sparse that there is much payoff to public or private sector R&D administrators from making greater use of them....R&D administrators do use ex post evaluations for political and organizational purposes, for example, to convince sponsors that they are interested in rational decision processes and that they are funding good work. However, the research evaluation literature between 1985-1990 contains very few demonstrations that evaluation makes any difference at all to critical decisions about the level and allocation of scarce scientific and technical resources (H. Averch, 1990, quoted in Kostoff, 1993, p. 175).

Although the task is formidable, we will attempt to outline a rationale and methodology for R&D evaluation, specifically for the federally funded, Department of Education R&D center, the Center for Research on Evaluation, Standards, Student Testing (CRESST). CRESST is the major federal R&D center developing better ways of assessing student performance, procedures sometimes referred to as "alternative assessment" or "authentic assessment" or "performance assessment." It is headquartered at UCLA with cooperating researchers at several universities (including Colorado, Stanford, and Pittsburgh) and other related institutions.

Dimensions of Evaluation

Two dimensions for evaluating R&D are merit and social worth. One can have good research which is meritorious on its own terms but which has little or no social impact, and one can have bad research which has quite a lot of impact. A recent example of the

latter is The Bell Curve by R. Herrnstein and C. Murray (1994), which has had tremendous short-term social impact but which probably could not be published in quality academic journals because it could not pass peer review (House and Haug, 1995). One might relate quality to impact in R&D something like this:

		Impact	
		High	Low
Quality	High	A	B
	Low	C	D

Ideally, one would want high quality research with high impact, or, if not, then low quality research with low impact or (reluctantly) high quality with low impact. The worse case scenario is low quality with high impact. No doubt, there are other important dimensions on which research can be judged as well, including its ethical character, which so far has been dealt with through university review boards (Howe and Dougherty, 1993; Sieber, 1982). Although the ethical/moral dimension is extremely important, here we limit ourselves to mostly impact, since that seems to be relatively undeveloped at this time (though not necessarily the most important).

Although there is not always agreement on the quality of particular pieces of research, there is some agreement on procedures for judging research quality, at least among researchers. The method of choice is peer review by leaders in the field, according to agreed criteria, e. g. consistency with previous research, conclusions consistent with data, appropriate data collection techniques, etc. (Chubin, 1994). CRESST has been

subjected to a number of peer reviews by its primary funding agency, the Department of Education, and has repeatedly been given high marks on the basis of its research proposals, in fact, beating out the competition. We will not spend time on the quality dimension, though there is quite a lot to be said about the adequacy of peer review procedures (Chubin, 1994). If one assumes that the R&D is high quality, what then?

Evaluating applied R&D is complicated in that it is presumed to have payoff beyond the research discipline. Hence, whether applied R&D is successful requires judgments of more than research experts in the field. How can one evaluate the impact of R&D efforts? There are few acceptable methodologies for doing so. In fact, from an evaluation perspective, one might compare this situation to 1965, the year of the Great Society legislation, when evaluation of large social and educational programs suddenly became a national priority, though no one knew how to do it. Many ideas were advanced and debated until a field of program evaluation gradually emerged after several years.

Shadish (1989) contends that concepts from program evaluation are useful for evaluating R&D as well. He distinguishes between internal-external and process-outcome evaluation criteria, noting that almost all methods for evaluating science use internal criteria (internal to the discipline itself), reflecting the reluctance of scientists to examine their work from the criterion of social worth. However, program, product, proposal, and personnel evaluation, from which the unified evaluation field has emerged, are

not the same as science or R&D evaluation. Though the evaluation field offers concepts, it does not provide clear direction.

The special problem with R&D impact evaluation is that the effects of R&D projects are so uncertain. Effects may take quite a long time to emerge and the pathway of R&D influence may be indirect, delayed, or even obscure. "The uncertainties associated with R&D, its multiple consequences, cumulative nature, and transferability all help to explain why the evaluation of R&D is so difficult....A critical problem in evaluating R&D activities is the long and uncertain time frame in which 'results' may be observed" (Melkers, 1993, p. 44). This pervasive uncertainty about "payoff" far down the line makes R&D impact evaluation difficult.

Anticipated vs. Unanticipated Impact

One way to approach this problem would be to separate potential R&D effects into the unanticipated and the anticipated. We might admit that ultimate long-term effects of R&D are beyond our present ability to assess for the most part, although one might be able to determine some unanticipated effects by retrospective case studies. R&D "spin-offs" might involve applications of ideas, products, or technologies that were not planned or anticipated at the beginning of the project (Brown and Wilson, 1993). These effects may occur by application of the idea to new markets or domains not anticipated, the classic case being military R&D adapted to civilian uses.

New technologies also may be reworked or combined with other ideas or technologies to produce "second-generation" technologies. In other words, new ideas and technologies differ

greatly in their "robustness" or ability to generate further (unanticipated) ideas down the line. These second-generation technologies often grow from the failures and learning experiences of the initial project. Sometimes they are linked to "enabling" technologies that solve critical technological or marketing problems. Although most spin-offs are accidental by-products, they significantly enhance the payoff from the original technology (Brown and Wilson, 1993).

On the other hand, there are R&D effects, impacts, or influences that are anticipated and foreseeable. That is, for particular types of R&D one might expect to find effects more likely with some groups rather than others, depending on the nature of the enterprise. In basic research in the physical sciences one might expect to find responses from colleagues at the theoretical level, but not expect to find marketable products. If this is so, then tracing influence by bibliometric techniques for specially-defined populations would make sense for basic research (ignoring problems with such techniques for the moment).

(We are using the terms "effects," "impacts," and "influences" somewhat interchangeably, although they carry different connotations. Impact is the favored term in Washington these days, perhaps because it reflects direct, simple, unidirectional, almost physical, effects and influences. "Influence" suggests something more tentative, effects probably mediated and modified through groups of people. "Effects" is the most general term and suggests that there may be far-reaching, unanticipated, and undetected influences.)

In a review of the dissemination literature Weiss (1989) recommended that one should distinguish between "knowledge utilization" and "innovation diffusion." The two are not the same. Most research does not produce innovations and most innovations are not derived from research. She suggested that one should define potential clients broadly, including policy makers and the public, that one should think in terms of intermediate organizations by which knowledge can be diffused, and that one should think less in terms of discrete studies than in terms of compilations of evidence on particular subjects.

One R&D impact evaluation, unusual for its size, consisted of 31 retrospective case studies of technology absorption and transfer for the New York State Energy Research and Development Authority (Kingsley, Bozeman, Coker, 1995). This retrospective "aggregate case approach" compared data across cases qualitatively and quantitatively. Two change processes emerged from these cases, those in which "technology absorption" into cooperating agencies occurred and those of "technology transfer," in which third parties removed from the development used the results for their own purposes. The two processes were different.

The cases were scored according to progress attained along these lines:

Technology absorption	*Technology transfer*
no impact	no impact
project impact	project impact
absorption	transfer object created
utilization	transfer strategy created

organization impact	transfer activity
	out-the-door
	utilization
	organization impact
benefit	benefit

Where no one used the technology at all, it was not for lack of trying on the part of the developers. Ordinarily, no use occurred because a key actor withdrew support for any of a number of reasons, including belief that the market wasn't there, that the technology wasn't good enough, or because of local politics. Many of these projects involved risky prototype development by several organizations and were sponsored by sole source funding rather than competitions. The projects encountered goal conflict, especially between academic and industry participants. Hence, evaluating effects solely in terms of transfer efforts can be misleading since considerable effort may lead to no effects ultimately.

Projects in which there was high absorption but low transfer often involved large complex process technologies. Only a few organizations were involved, with subcontractors taking the lead. Funding was from one or two sponsors, and transfer efforts were minimal. By contrast, successful market-induced transfer to third parties involved many participants (4 to 9), with a public agency playing a large part in the impetus. Sponsor and contractor-induced transfers usually involved an active public sector agency which created a market through regulation and encouraged networks of suppliers and vendors for other public organizations. Often the technology consisted of knowledge products, e. g., computer

software. For the most part, absorption was more robust than transfer. Transfer was successful when public sector actors were the targets and consisted many activities rather than a uniform process.

Bozeman, Papadakis, and Coker (1995) also studied relationships between federal R&D (hard science) labs and commercial corporations. These federal labs were free-standing, government-sponsored enterprises working directly with companies to develop products. Where such relationships obtained, the companies were satisfied if a product was developed. On the other hand, the creation of new jobs, a major rationale for such cooperation, was nil. There could be successful product development without job creation.

Methods for Evaluating R&D Impact

In the sparse literature on evaluating R&D enterprises, here are the major data collection methods (omitting patent analysis as not relevant):

Return on investment--Used in business and economic analyses sometimes. One needs to calculate benefits and what they cost to make this evaluation work (Link, 1993). There are cases where precise estimates are not needed, e. g., a project is costly but has little or no payoff, but these methods have not been applied with much success in R&D evaluation for the most part, even though this is the kind of information legislators would like to see. (In fact, Bozeman, Papadakis, and Coker, 1995, p. 44, described the current R&D evaluation environment as one of "desperately seeking numbers.")

Although it is easy to imagine such an evaluation, e. g., put an R&D product into place and measure the increase in achievement test scores related to costs, these measurements are beyond our evaluation capabilities currently. In fact, this type of analysis was the intent of the Follow Through program, with all its attendant difficulties. The dozen or so Follow Through early childhood education "models" produced so many different results that there was as much variance of student achievement within each model as between models. In other words, the same model at different sites produced quite different results. Such contextually sensitive results does not provide a stable basis for assigning costs compared across sites.

In fact, defining production functions for education has not been successful generally (Monk, 1992). There are several reasons for this, the most likely being that educational production (achievement by students) is caused by an array of interactive factors so that introducing a new educational technology leads to different results, depending on what the factors are in a given context. One can evaluate success within a context but generalizing across contexts is much more difficult, as has been demonstrated repeatedly in the evaluation literature.

This disappointing result has not been limited to educational programs. In a review of US federal R&D evaluation (hard science and technology), Kostoff concluded, "Cost benefit analysis has limited accuracy when applied to basic research because of the quality of both the cost and benefit data due to the large uncertainties characteristic of the research process, as well as

selection of a credible origin of time for the computations" (Kostoff, 1993, p. 174). R&D evaluation has relied heavily on peer review procedures instead.

Not even business firms heavily engaged in R&D use quantitative measures to evaluate the productivity of their operations. In one study only 20% of the leading 34 R&D firms used any kind of quantitative productivity measures; 59% used none at all. And there was skepticism about doing so. One R&D firm director said, "...attempts to quantify benefits of R&D have led to monstrosities that caused more harm than good" (Roessner, 1993, p. 188). Barring a significant reconceptualization, this type of evaluation doesn't look promising at the moment.

Bibliometric methods, such as ascertaining the number of publications and citations (as well as co-citation analysis, co-word analysis, scientific mapping, and patent citations) for authors or groups of authors, might be useful for judging the success of more basic research (assuming a continuum from basic to applied). Number of publications is an indicator of productivity or output more than quality, while citations may reflect quality to some degree. However, bibliometric experts caution against using citations to judge individual scholars because the margin of error is too great. Citations are better applied comparatively to similar groups. Citations may also be more indicators of use than quality (e. g., Toffler's Future Shock was the most frequently cited scientific publication over one time period but is unlikely to influence scientific research very much).

However, Shadish (1989) notes that publication and citation indices are perhaps the only widely accepted indicators of scientific quality. Presumably, seminal ideas, authors, and research units would show up in publication and citation lists. He found that highly cited works are judged to be of high quality but that most high quality works are not cited frequently. Awards, honors, and research grants received can add to a mixed list of indicators. However, one must be cautious about the consequences of adhering strictly to such indicators (Shadish, 1989).

Currently, a significant portion of research funding for British universities is based on bibliometric information (Johnes and Taylor, 1990; Melkers, 1993). Economists investigating the British research selectivity evaluation arrived at the unsurprising conclusion that research output depends very heavily on input of resources, and that output could not be fairly assessed without reference to input. There are also many sources of bias in citation studies, not the least of which is that badly conducted studies are sometimes singled out for criticism (and citation).

For applied R&D bibliometric indicators seem insufficient. For example, the National Institute for Occupational Health in Sweden, a research institute of medical researchers, physicists, engineers, and psychologists, has the mission of furthering the occupational health of Swedish workers. Their current practice is to award the highest number of "merit points" to refereed articles in international journals and then allocate internal funds on that basis. Such procedures directs the institution away from its purpose of helping workers. Although publications and citations are relevant for

judging some work, one would expect more concrete payoff from applied research.

Surveys are useful for tracing influence in large populations, assuming that the type of influence is not too complex to be recorded in a survey instrument or brief interview. For example, Stalford and Stern (1990) carried out a survey of school districts in the US, asking whether the district superintendents had heard of products from the US Department of Education regional labs, R&D centers (such as CRESST), and the ERIC system, and if they had, what they had done with the information. The survey covered a large sample of potential users of R&D products, i. e., school districts. The limitation of the study was that "influence" was expressed in simple terms on one page.

In a study of government science and technology evaluation in the Canadian government by the Office of Comptroller General, 90 percent of the evaluations used client or stakeholder surveys, 40 percent used surveys of expert opinion, 67 percent used literature reviews, 26 percent case studies, and 13 percent non-client interviews. Only one study used citation analysis; none used peer review. According to the study, successful client surveys included experts in the survey process, segmented survey respondents, and used experienced interviewers well-versed in the subject to probe the issues

Case studies are better at tracing complex events, including the most diffuse and complex outcomes, but they have problems of subjectivity, sampling, and comparability (Stake and Easley, 1978; Kingsley, 1993; Stake, 1995). Also, they are expensive to conduct so

that their number must be limited. However, there probably is no better way to assess complex influences over a period of time. Of course, even if one employs case studies, the problem of how long it takes for R&D to register effects remains as much a problem as before. Case studies can provide finely textured portrayals of effects but are no less confined to limited time periods.

Where case studies have been used in R&D evaluation, they have been effective in generating specific information and giving managers a "feel" for the programs. In general, "soft" techniques have been deemed more appropriate because of the abstract nature of R&D programs (Barbarie, 1993). One might note that, "In business as in government, evaluation of basic research is recognized as necessarily a judgmental process, best accomplished through use of informal, largely qualitative methods" (Roessner, 1993, p.199).

The New York State Energy case studies also offer an intriguing methodology, though its implementation would be extremely expensive. Multiple case studies (31) provided a broad view of the effects of the New York program over time (though the purpose of the research was to ascertain characteristics of successful projects). The two influence pathways of "absorption" and "transfer" and their contingencies are suggestive of influence that might be expected for other R&D.

Network analysis is another possibility, though not one that has been tried. If one conceptualizes influence as being exercised through personal networks of people and contacts, one can apply methods for analyzing the networks themselves, including intermediate pathways through which R&D ideas and products might

exercise influence, and hence anticipate success as a function of the networks in advance of utilization, thus circumventing the time lag to some degree.

For example, there is extensive evidence in the innovation diffusion literature that face-to-face interaction is critical to the diffusion and acceptance of innovations, a large research literature produced by quantitative geographers, rural sociologists, and medical and educational researchers (Hägerstrand, 1967; Rogers and Shoemaker, 1971, Havelock, 1971; House, 1974). Similar findings have emerged in the technology diffusion literature, with one study claiming that national research performance stems not only from the activities of individuals but also from their interactions, with some countries having distinct advantages in this regard (Dalpe and Anderson, 1993).

Of course, R&D networks can consist of many things: connections within R&D facilities that link research and development; connections among potential adopters or users of R&D materials; connections among R&D organizations themselves; connections among members of user organizations; connections between journals and their readership; invisible colleges; newspaper readership and community influence organizations, and so. Hypothetically at least, one could define networks of influence for particular projects. The more R&D ideas are diffused to the general public (as in AIDS research), the more these networks might function as mass communication, as in electoral campaigns. The more the technology requires extensive changes in behavior, the more the diffusion might resemble classic technological change.

The advantage of conceptualizing R&D influence as exercised through personal networks is that the process could be subjected to precise analyses, even mathematical analyses, that can be applied to networks generally, whether these networks are based on face-to-face contacts or attitude change through mass communication (Hägerstrand, 1967; House and Long, 1974; Hood, 1989; Mill and Stephens, 1992; Zaller, 1992). Using the concept of the network turns the unknown into something more familiar and analyzable. No doubt, such a conceptualization of the problem would not account for all influence, but might be useful.

In summary, our ideas for evaluating R&D impact are to divide the effects into anticipated and unanticipated and to focus on the anticipated. One might anticipate that R&D work will have effects, influence, and impact on certain identifiable populations and that some influence will be carried through networks of people engaged in face-to-face interaction. Other influence might be carried through impersonal networks, such as the media. Hence, one could define certain network populations and trace influence, and possibly define characteristics of networks that would anticipate and facilitate success in advance of ultimate effects. For example, one might predict that a particular network would be highly unlikely to diffuse the R&D idea advanced because of its structure.

The Case of CRESST

CRESST is not a tightly-organized, closely-managed, single-dimensioned entity. It is a consortiuum of well-known scholars who cooperate voluntarily and largely on their own terms. The activities are many, diffuse, and often autonomously initiated by the scholars

themselves. There is no reason to believe that such a complex enterprise will be any less productive scientifically than a tightly managed one. A diverse structure poses a problem for evaluators, however, in that effects may be generated from a number of sources to many different groups.

The official CRESST program structure for researching and developing "alternative" modes of student assessment consists of the following:

Program One: Building the infrastructure for improved assessment

Project 1.1: Synthesis and collaboration

Project 1.2: Technical assistance and dissemination

Program Two: Designing improved learning-based assessments: Prototypes and Models

Project 2.1: Designs for assessing individual and group problem solving

Project 2.2: Deep understanding of content knowledge

Project 2.3: Complex performance assessments: Expanding the scope and approaches to assessment;

Project 2.4: Quantitative models to monitor the status and progress of learning and performance and their antecedents;

Project 2.5: Analytical models for performance and delivery standards.

Program Three: Collaborative Development and Improvement of Assessments in Practice

Project 3.1: Studies in improving classroom and local assessments

Project 3.2.: State accountability models in action

Project 3.3: Policy and cost studies

Of course, programs and projects written into funding proposals often do not resemble the actual operations very closely, especially in consortia of major institutions, each part of which tends to function with considerable autonomy. If one took the structure and operation of these paper programs literally, one would be misled in some cases. One must look at operations rather than only proposals. More specifically, CRESST administrators envision their categories of work along these lines (Dietel/Herman memo, Feb. 2, 1995):

A. Assessing assessment: assuring the quality and validity of assessment systems (e. g., the CRESST criteria);

B. Effective models for developing alternative assessments;

C. Systems and implementation strategies that serve accountability and improvement;

D. Approaches to assuring technical quality;

E. Technical methodologies to optimize cost effectiveness;

F. Models for standards setting;

G. Models for assessing and implementing opportunity to learn; assuring equity in assessment;

H. Effective reporting formats;

I. Effective professional development models;

J. Costs and effects of large-scale assessment;

K. Lessons in state assessment policy.

The purpose of CRESST is to develop new and better ways of student testing, especially focused on "alternative assessment." One

might expect the effects of such R&D work to be registered closer to educational practice than theoretical work might be, though some CRESST research is certainly theoretical. Again, one might divide potential influence into the anticipated and unanticipated. Anticipated effects would include influence on groups and individuals with whom the Center is actively working and targeting, including,

--those who attend conferences sponsored by CRESST;

--those who order products or other information from CRESST;

--those who work directly with researchers and developers;

--readers of journals in which the ideas are published;

--other R&D specialists in the same fields;

--targeted "user" populations, such as teachers, school districts, state education agencies, and test developers;

--the public at large, or some segment of it.

One might expect influences to be registered differently within these groups. Ron Dietel, director of communication at CRESST, defines the major CRESST constituencies this way, which concurs closely with our own analysis above:

- Researchers
- State and District test directors (including Title 1 directors)
- Practitioners
- Test developers (commercial and otherwise)
- Foundations
- Policy makers
- Media

These constituency groups can be ordered as to the type and extent of influence one might expect to occur. One would expect policy makers, test makers, and the public to react in different ways to CRESST R&D than do researchers. Some groups operate closer to the R&D work and with the original concepts, products, and researchers, while other groups are far removed and only distant potential users, such as the vast majority of teachers, administrators, and the public. Nonetheless, the teachers and students are presumably the ones ultimately affected by these efforts.

Another way of characterizing these relationships is that researchers and developers exert their influence through established networks, such as those defined by journals or conferences, or by trying to establish new networks, such as by building enduring contacts with those who want to work with the ideas. Hence, one might characterize influence activities partly as network building and/or utilization. The advantage of conceptualizing R&D influence this way is that one can investigate anticipated pathways of influence by thinking of network construction, thus introducing an intermediate set of constructs that can be analyzed for potential influence, i. e., the characteristics of the networks themselves.

According to findings from the New York State Energy studies (admittedly different types of R&D activities), one might add relationships and networks among the sponsors (the US Department of Education, plus others), the contracting agency and its subcontractors (UCLA, the participating universities), and other participating organizations (such as school districts, commercial

test organizations, and professional organizations). According to the New York study, certain contingency relationships among these groups are related to eventual absorption and transfer.

Another complication is the unit of influence. Should it be the persons doing the R&D, the project or program, or the ideas, concepts, and materials that are the focus of study? The answer would seem to be, "It depends." In the early days of R&D it would make sense to trace the influence of particular persons. The originators' personal influence is likely to be registered more at the beginning. If there are several people working on the same ideas, the program or project would seem to be an appropriate unit of analysis. However, sometimes it is not clear where one project stops and another begins. Programs overlap and different ideas and projects are fitted together as an administrative (and funding) convenience because sponsoring agencies think (and report) in terms of projects and programs.

As ideas and products spread, however, contact with the originators or even the projects may become less likely so that the ideas, concepts, and products themselves, e. g., "alternative assessment" in a vague sense, may be the way change is registered, without reference to the sources. By a later time influence may be spread through so many intermediaries that its sources may not be known (or deliberately ignored). The ideas may also be transformed and renamed, and/or attributed to incorrect origins, a likely event in that entrepreneurial agents are often instrumental in later stages of influence. In fact, the manner in which educational research findings are used by policy makers, use based on third or fourth hand

information or media accounts, is similar to the formation of mass opinions (Weiss, 1989; Zaller, 1992).

For example, perhaps the most successful concept in educational R&D over the past few decades is meta-analysis (Glass, Smith, McGaw), the use of which has spread far beyond its original uses in education. Medical research in particular has made extensive use of meta-analysis, but it is often the case that the technique is claimed to have been developed by researchers in medicine without attribution to correct the original sources. Such obfuscating screens are not unusual in the competitive world of R&D.

Applying Concepts and Methods to CRESST

How then have we decided to evaluate the impact of CRESST? We divided the potential influence/impact groups into the following:

Researchers--To assess the impact on researchers, we conducted several analyses of publications, including numbers of publications of major CRESST researchers, numbers of citations, most important publications (in the judgment of CRESST), how these citations were used, and what kinds of journals these citations appeared in. We used the Social Science Citation Index as our main data source, but also counted publications in important educational journals not included in the citation index, such as the Educational Researcher. This moves somewhat beyond regular citation analysis.

Test directors--We surveyed members of the national association of test directors through the mail, asking them what influence the CRESST products and ideas had on their behavior and domain. We conceived test directors to be an important audience

because of their gatekeeper function regarding testing, even though many had little contact with CRESST.

Practitioners--Practitioners consist of administrators and teachers. School superintendents have been surveyed about their recognition of R&D products in general (Stalford and Stern, 1990), and though they are an important group, it is not likely many are well informed about alternative assessment at this time. The same is true for principals. The exception would be administrators whose districts and schools have participated directly in CRESST activities.

In the Stalford and Stern study, recognition of federal R&D institutions was high (over 90 percent). Of the 64% of the districts recognizing the federal R&D centers, 52% received products, with the majority of these (37%) reporting frequent use. However, when asked about "particularly useful" materials the R&D centers received only 2% mention. In general, the Department of Education evaluators were surprised by the positive responses. In another study the federal R&D centers and laboratories were judged not effective in disseminating their work (Center for Leadership Development, 1984).

We conducted two studies to trace the influence of two of CRESST's most successful products, as judged by CRESST administrators. Telephone interviews were conducted with people nominated by the developers, then with people nominated by the interviewees, and so on, until a chain of influence was traced from developers to use in the field, as far as that was possible. The two

products whose influenced we tried to follow were nominated by CRESST as two of their most influential products.

Direct contacts--CRESST maintains a list of people with whom it has had contacts of one sort or another, consisting of about 12,000 entries at this time. A brief, one-page mailed survey was sent to a random, stratified sample to ask them if and how they had used CRESST products and ideas. Most people had contact by attending CRESST conferences or by soliciting CRESST material through the mail.

Teachers--Usually, teachers are considered the ultimate audience for CRESST impact, yet few projects have worked directly with teachers. One might expect that it would take quite a long time for a substantial number of teachers to use CRESST products. CRESST research reports were surveyed to gain an idea of how teachers will react to notions of alternative assessment, plus brief interviews were held with a few teachers who had tried to develop alternative assessments for their classrooms. Teachers also composed a subgroup in the survey of direct CRESST contacts.

Policy Makers--Within this data collection, there are some important things omitted. First is the potential impact on policy makers. Certain CRESST people worked with policy makers to set testing policies at the state and national levels. This is an important influence, even though it occurs through individual behaviors. However, when we solicited from CRESST researchers what influences they thought they had, the nominations were highly erratic with regard to policy makers. Most did not name any policy makers, while other mentions were uneven. Hence, we omitted this

potentially important influence because we could not see how to approach it systematically within our resource constraints. Even if we had been able to nominate potential policy influences, it seemed unlikely we could survey policy makers with mailed questionnaires or through telephones. This is a significant limitation of the evaluation. We have done a rough content analysis of the CRESST nominations, however.

The public--One might expect CRESST to inform the public as well as professionals about alternative assessment. The Nexes data base was searched to ascertain how many times alternative assessment appeared as a topic in newspaper articles, even when the newspaper did not mention CRESST by name. One cannot attribute mentions of alternative assessment to CRESST influence by any means, but no doubt there is an indirect influence in many cases.

Case Study--We also considered one or two overall case studies of some scale. One can have any number of pieces but these pieces may fit together in unexpected ways because politicians, tests, teachers, etc., interact with each other in powerful ways. The ultimate challenge is to see how everything works together in a school district or state. The principal investigator volunteered to do a case study of Hawaii in December, but the idea was greeted with a lack of enthusiasm by CRESST administrators. This is an omission we find particularly distressing.

Network Studies. Second, we did not conduct any empirical studies along the networking lines suggested earlier, again mostly because of lack of resources. One might define important potential

networks on which the new ideas and products might travel and assess the degree to which these networks had been penetrated, thus enabling an estimate of the impact of the R&D before the actual influence had occurred. For example, one might define the test director network as a whole and estimate how far CRESST ideas had traveled within this social structure. However, such an analysis would require far more effort than we were able to muster.

By assessing the impact of CRESST products and ideas on several groups, one can make a judgment about the overall CRESST influence. Of course, putting all these disparate studies and pieces of information together to arrive at an overall judgment is no easy task. That is the last task of this report, and it presents several data synthesis problems of its own, apart from the separate studies. On the other hand, it is difficult to see how any one study could accurately assess the impact of all the CRESST activities.

In summary, one might imagine an R&D technology ("alternative assessment") sweeping the country from its center of creation to transform education. However, that is not how R&D influence and impact seem to occur. Rather, the R&D center works with diverse groups of audiences simultaneously (and opportunistically) to register different kinds of influence within each group. If development and diffusion of the technology across the nation is the grand, idealized picture, then this evaluation is limited to penciling in separate and discrete segments of the picture so that one has an idea of its shape and extent. The evaluation arrives at a calculated judgment about R&D influence based on disparate pieces of evidence, rather than a complete account.

References

Averch, H. A. (1993). Criteria for evaluating research projects and portfolios. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 263-277.

Barbarie, A. J. (1993). Evaluating federal R&D in Canada. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 155-162.

Bozeman, B. (1993). Peer review and evaluation of R&D impacts. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 79-98.

Bozeman, B., Papadakis, M., Coker, K. (1995). Industry perspectives on commercial interactions with federal laboratories. Atlanta, GA: School of Public Policy, Georgian Institute of Technology. 75 pages.

Brown, M. A. and Wilson, C. R. (1993). The temporal dimension of R&D evaluation: Incorporating spin-off benefits. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 244-262.

Center for Leadership Development. (1984). Creating and disseminating knowledge for educational reform: Policy management of Education's regional educational laboratories and national research and development centers. A report to the National Council on Educational Research. Los Angeles, CA. 116 pages.

Chubin, D. E. (1994). Grants peer review in theory and practice. Evaluation Review, Vol. 18, 1, 20-30.

Cozzens, S. E. (1993). US. evaluation of strategic research: Closing the gap with Europe and Asia. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 229-244.

Dalpe, R. and Anderson, F. (1993). Evaluating the industrial relevance of public R&D laboratories. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. (207-228).

Hägerstrand, T. (1967). Innovation diffusion as a spatial process. Chicago, IL: University of Chicago Press.

Havelock, R. G. (1971). Planning for innovation. Ann Arbor, MI: Center for Utilization of Scientific Knowledge, Institute for Social Research.

Herrnstein, R. J. and Murray, C. (1994). The Bell Curve. New York: Free Press.

Hood, P. D. (1989). How can studies of information consumers be used to improve the education communication system? Paper presented to the American Educational Research Association, March 29, San Francisco, CA.

House, E. R. (1974). The politics of educational innovation. Berkeley, CA: McCutchan Press.

House, E. R. and Haug, C. (1995). Riding the bell curve. Educational Evaluation and Policy Analysis. Summer.

House, E. R. and Long, J. M. (1974). Applying directed graph theory to faculty contact structure. Paper presented at American Educational Research Association, Chicago, IL, April 19.

Howe, K. R. and Dougherty, K C. (1993). Ethics, institutional review boards, and the changing face of educational research. Educational Researcher, 22, 9, 16-21.

Johnes, J. and Taylor, J. (1990). Performance indicators in higher education. Buckingham UK: Open University Press.

Kingsley, G. (1993). The use of case studies in R&D impact evaluation. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers, 17-42.

Kingsley, G., Bozeman, B., Coker, K. (1994). Technology transfer and technology absorption: An aggregate case approach to evaluating RD&D impacts. Syracuse, NY: Center for Technology and Information Policy. Syracuse University. 61 pages.

Kostoff, R. (1993). Evaluating federal R&D in the United States. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 163-178.

Leslie, S. W. (1993). The Cold War and American science. New York: Columbia University Press.

Link, A. N. (1993). Methods of evaluating the return on r & d investments. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 1-16.

Melkers, J. (1993). Bibliometrics as a tool for analysis of R&D impacts. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers, 43-61.

Mills, Stephan R. and Stephens, K. Gwen. (1992). Impact of education dissemination practices on California schools. Far West Lab, San Francisco, CA. ERIC Document 354878.

Monk, D. H. (1992). Education productivity research: An update and assessment of its role in education finance reform. Educational Evaluation and Policy Analysis, 14: 4, 307-332.

Roessner, D. (1993). Use of quantitative methods to support research decisions in business and government. In B. Bozeman and J. Melkers (Eds). Evaluating R&D impacts: Methods and practice. Boston MA: Kluwer Publishers. 179-205.

Rogers, E. M. and Shoemaker, F. E. (1971). Communication of innovations. New York, NY: Free Press.

Shadish, W. (1989). Science evaluation: a glossary of possible contents. Social Epistemology, 3, 189-202.

Shadish, W. R. (1989). The perception and evaluation of quality in science. In B. Ghoulson, W. R. Shadish, Jr. R. A. Niemeyer, A. C. Houts. (Eds.). Psychology of science: Contributions to metascience. New York: Cambridge University Press. 383-426.

Sieber, J. E. (Ed.). (1982). The ethics of social research. 2 vol. New York: Springer-Verlag.

Stake, R. E. and Easley, J. A., Jr. (Eds.). (1978). Case studies in science education. Center for Instructional Research and Curriculum Evaluation. Urbana, IL: University of Illinois.

Stake, R. E. (1995). The art of case study research. Thousand Oaks, CA: Sage Publications.

Stalford, C., and Stern, J. (1990). Major results of a survey on the use of educational R&D resources by school districts. American Educational Research Association, Boston, MA, April 16-20. ERIC Document 329 212.

Wargo, M. J. (1994). A congressional agency's evolving evaluation approach to research and development. Paper presented to American Evaluation Annual Meeting, Boston, Nov. 4, 1994.

Wargo, M. J. (1994). Impact of the President's Reinvention plan on federal evaluation activity. Paper presented to American Evaluation Annual Meeting, Boston, Nov. 3, 1994.

Weiss, C. H. (1989). Improving the dissemination of research, statistical data, and evaluation results in education. Cambridge, MA: Harvard University. Mimeo.

Zaller, J. R. (1992). The nature and origins of mass opinion. Cambridge, UK: Cambridge University Press.

Chapter 2

Beyond Counting "Pubs:" An Improved Bibliometric Method for Evaluating CRESST's R&D Impact.

Scott F. Marion
University of Colorado, Boulder

Abstract. This study focused on one component of a multi-method evaluation of CRESST: its impact on the measurement research community. Several bibliometric techniques were employed to evaluate the impact of CRESST on the research community, including traditional bibliometric methods such as publication and citation counts. Also, we have attempted go beyond these techniques to determine how CRESST research is used by other researchers and how CRESST is shaping the alternative assessment research agenda in the measurement community.

These analyses indicate that CRESST partners are quite productive and publish their works in high status outlets. Although many bibliometricians dismiss publication counts because of lack of information about research quality (compared to quantity), the peer review process associated with the highest status journals seems to be a reasonable indicator of quality. We conclude that most CRESST research is of fairly high quality as determined by peer review.

The core group of eighteen CRESST partners produced 90 articles, books, book chapters, and technical reports that were cited by other researchers. These 90 articles were cited 424 times since 1990. Most reports were published in 1991 or later. A substantial portion of these citations appeared in articles written by other CRESST partners (24%) or by the author of the cited works, "self-citation" (17%). However, a

majority (59%) appeared in articles written by researchers not associated with CRESST.

The number of citations per article is not normally distributed. Rather the majority of articles did not receive any citations, and most received only one or two. For example, of the 81 citations for eleven articles for Robert Linn, 46 were for a single article (Linn, Baker, & Dunbar, 1991). The work of CRESST partners is cited at a high rate with an average of 4.7 citations per cited article. Discounting the self-citations resulted in an average of almost 4 citations per article.

If CRESST research were cited only in "low status" journals, one might infer that the measurement community did not think it was important. Conversely, if CRESST work is cited in high status journals, one might infer that CRESST work is considered important. CRESST articles were cited in journals with an average status rating of 2.3 (of a possible 3). Some lower ratings were an artifact of the rating procedure such that a widely cited article might "lose" points because of citations in practitioner journals. The high average rating of citation sources is evidence that CRESST work is considered important to other researchers.

Relying on an addition to bibliometric methods, we focused on fourteen key CRESST articles to determine how these articles were used by other researchers. All except two of these nominated articles were cited at least six times. In total, these fourteen articles were cited 175 times. On a three point scale of centrality of "use," few citations received a "3." Most citations merited a rating of "1" or "2." Use ratings approaching 2.0 might be considered an indicator of a very "useful" article. A finding obscured by these analyses is the difference in the way empirical and conceptual articles were used. Articles reporting new

empirical findings or summaries were cited to support a specific point and be rated "2." Conceptual articles, such as Linn, et al. (1991), have more "3" ratings than empirical ones, but also more "1" ratings. The high use ratings for the fourteen articles demonstrate that other researchers consider CRESST research important.

To determine the boundaries and extent of CRESST's influence, we also conducted an ERIC search for the term "performance assessment," limited to journal publications in 1994 and 1995. There were 54 performance assessment articles published during this 15 month period, 10 of which were authored by CRESST. Having just under 19% of the articles published in this field by a single research program is an indication of influence. Twenty percent might be considered a critical mass by most standards.

To discover how often CRESST work had been cited in this sample, we searched the reference lists of 35 (of the 54) articles we could locate. CRESST research was cited 90 times in the nine (of 10) CRESST publications we located and 42 times in the 26 non-CRESST articles. All CRESST articles included references to other CRESST research, except one. Of the 26 non-CRESST articles, 12 did not include any post-1990 CRESST citations. The 42 citations from these articles came from 14 sources.

Among the CRESST sources, one article, Baker, O'Neil, & Linn (1994) accounted for 28 citations, far more than any others. If this outlier is removed, the remaining seven CRESST articles had an average of approximately nine CRESST citations per article, more than five times the amount of citations per non-CRESST authored article. This is indirect evdience that CRESST researchers form a school of thought to some degree. Nonetheless, CRESST was still well represented by non-CRESST

authors, and most CRESST citations were in the high status journals. These ERIC analyses present evidence that CRESST authors contributed a critical mass of publications and CRESST research was cited at a high rate by non-CRESST researchers.

Taken together, the publications, citations, status and use ratings, and ERIC analyses all point to CRESST's impact on alternative assessment research and offer persuasive evidence that CRESST researchers are strongly influencing the research agenda within the measurement community.

CRESST Background

The purpose of 1990 funding award to CRESST was to improve assessment at all levels of education. CRESST's major goals for their research and development program are these:

1. Provide leadership to improve policy and practice at the national, state, and local levels.

2. Develop theories and models to improve the quality of student performance measures.

3. Develop new theories and models for understanding and assessing the quality of schooling.

4. Clarify the role of assessments in improving educational practice.

5. Improve the understanding of assessment policy and its contribution to educational improvement (CRESST, 1992 p. 2).

CRESST attempted to reach these goals through three research programs:

Program One: Building the Infrastructure for Improved Assessment...exists to strengthen the infrastructure for improving assessment practices across the country and to assure CRESST's impact on education policy and practice (CRESST, 1992p. 10).

This program fosters collaboration among researchers working on similar problems and between researchers and policymakers at all levels.

Program Two: Designing Improved Learning-Based Assessments: Prototypes and Models . Program Two creates new designs for assessing student performance and new models for analyzing and validating assessment results (CRESST, 1992p. 13).

This program is focused on technical and theoretical aspects of assessment design and implementation. While all three programs share the goal of influencing policy and practice, Program Two appears to target the research community.

> Program Three: Collaborative Development and Improvement of Assessments in Practice...is directed to questions of implementation and impact of assessment in policy and practice (p. 21).

Participants conducting research in this program are expected to work with teachers and policy makers to "develop strategies for facilitating the implementation of improved assessment methods, and to make recommendations for policy formulation (p. 21)."

This part of the evaluation is focused on Program Two, CRESST's influence on the measurement research community. The evaluation of a research program's impact on other researchers is difficult, though perhaps not as difficult as evaluating impact on practice and policy because the research community is smaller and more bounded than the practitioner or policy communities. Furthermore, there are existing methods for evaluating impact on other researchers.

Bibliometric techniques for counting the number of articles published and the number of times an article is cited are the most widely used methods for ranking scientists, research programs, and institutions. Many rankings of universities (e.g., U.S. News and World Report) are based on bibliometrics. Article counts are an important consideration in weighing the contribution of scientists in a given field and a key factor in determining tenure and promotion of faculty members.

Recent studies have indicated that counting the number of publications might not prove entirely useful because few scientific papers are ever cited. In a review of the top 4500 (of approximately 74, 000) science journals, only 45% of the papers published between 1981 and 1985 received one citation. Further, bibliometricians believe that approximately 10% of the journals receive 90% of the citations, so the 45% figure is likely an overestimate (Hamilton, 1990).

Using citations instead of number of publications can allow one to obtain a "weighted measure" of research output (OTA, 1991). One technique to compare institutions or different scientific fields has been used with some favor recently. The average number of citations per cited paper provides a method of sifting through the un-cited publications and focusing only on those likely to have an impact on the field (OTA, 1991).

In spite of its appeal there are drawbacks to bibliometrics. For ranking individuals across fields, it is important to keep in mind that citations are not made in systematic ways across or even within fields. Similarly, authorship is not allocated in consistent ways. Therefore, when bibliometric techniques are used, the citation practices of the respective fields need be compared. Within fields, especially if the field is small enough to have uniform citation practices, bibliometric procedures can be a useful indicator of impact.

Another major criticism is that bibliometricians rarely relate quantitative indicators to social and economic impact (Averch, 1990). In a sense, citations have become reified as something of unique worth, instead of a criterion based on "real world" import. While bibliometric indices might correlate with an external criterion, this link cannot be

assumed. Most science policy analysts recommend using bibliometrics in addition to other techniques to certify the social worth of research.

In this study we relied on bibliometric techniques to evaluate the impact of CRESST on the research community, but took heed of Averch's advice and included additional methods to certify the impact[1]. Most methods are not useful for evaluating the influence of one research program on a community of researchers. Counting publications might only reveal that the researchers who received the funding were the most productive. Citation counts are a step closer to evaluating impact but counting the number of citations is some distance from determining the impact of a research program on a community of researchers. In this study we used traditional bibliometric techniques, but have gone a step beyond to consider how CRESST research is used by other researchers.

Methods of this Study

First, we examined the entire list of publications and presentations of CRESST partners in order to estimate research productivity, then we performed a citation analysis to help us understand the impact of CRESST research on other researchers, and, last, we compared the citation and publication rates of CRESST researchers to the entire field of measurement researchers.

The first analysis was relatively straightforward. We scanned vitae from all CRESST partners (CRESST, 1995) to perform a count of standards and assessment related research and rated the research outlet in order to estimate its status or "influence" (Ciba Foundation, 1989). We counted

---

[1] Worth is often established through peer review and we have been working under the assumption that the quality or worth of this research is fairly high. However, other components of this evaluation focus on the social impact of CRESST's research.

all assessment (defined broadly) first-authored publications from 1990 through 1995. We recognize that articles published in 1990 or even 1991 were likely in progress prior to the 1990 funding of CRESST, but we used this cut-off date to ensure that we would be as fair as possible to CRESST. Further, by conducting the evaluation now, we are not counting articles resulting from CRESST research published in 1996 or 1997.

The following scale was used to rate the "influence" of each specific publication outlet:

> 3 = Highest status and influence journals. All AERA and NCME publications (e.g., *American Educational Research Journal, Educational Researcher, Journal of Educational Measurement, Educational Measurement: Issues and Practice*), other high status journals in education (e.g., *Harvard Educational Review, American Journal of Education*), and high status/influence journals in related fields such as psychology (e.g. *American Psychologist, Educational Psychologist*).
>
> 2 = Highly influential practitioner journals, but without the same status/influence on researchers as first tier research journals (e.g., *Educational Leadership, Phi Delta Kappan*). This category also includes second tier research journals which are thought of as important journals without the same status and influence as first tier journals or those with a more limited audience. (e.g., *Education and Urban Society, Journal of Educational Research, Educational Assessment*).

1 = Lower status research and practitioner outlets and technical reports. This category included research and other journals with little status or with such a narrow audience they would not have very much influence (e.g., *Education*, *Journal of Educating Computing Research*). This category also included less influential practitioner journals than those rated '2' above (see Appendix A).

Although it appears that combining research and practitioner journals in categories 2 and 1 above might be like combining apples and oranges, the important criterion for this study is influence and, in many cases, such practitioner journals as Phi Delta Kappan can have as much influence on other researchers as research journals.

Journal articles, books, book chapters, and certain technical reports and monographs were counted as publications. Technical reports only were counted if they were not followed by a journal article very similar to the technical report and if they were for a major institution such as RAND, CRESST, or ETS. Not all of the CRESST researchers listed technical reports on their vita. They were more common for CRESST partners involved with "think tanks," such as RAND. While such conferences as AERA and NCME are important for "spreading the word" about CRESST-related research, we did not include these outlets in this section because most CRESST partners could present at conferences as often as they wanted. CRESST researchers frequently are asked to participate in panel discussions or other types of sessions where they might not prepare original papers. We did not feel that we could adequately judge whether the presentation was an original paper or simply a discussion of previously articulated ideas.

The second analysis was more complex. Social Sciences Citation Index (SSCI) collects and catalogues citations from articles published in journals indexed by SSCI. The computerized version of SSCI allowed us to search for specific authors' works published after 1990 (through March 1995), when CRESST funding could have first been first expected to have yield results. The full list of researchers who have received CRESST funding is extensive. Therefore, we have limited the search to researchers forming a core group of CRESST measurement researchers (see Appendix B). We are certain that we have omitted some important researchers, but we feel that this purposeful sample adequately represents CRESST's influence on the measurement community. We used these search results to generate a list of articles where CRESST authors have been cited (see Appendix C). We used this list to summarize the citation counts, average number of citations per article, and status ratings (according to the scale above) to describe the results of the citation results.

These data were summarized for each CRESST researcher as follows: First, the total number of articles cited and the number of citations were counted, and then the average number of citations per article cited and the average influence per citation were calculated. As we did the analyses, we noticed that many CRESST articles were cited by other CRESST partners. Therefore, we included separate analyses for citations from other CRESST partners and self citations.

While the second phase of our bibliometric analyses, described above, is more extensive than most citation analyses, we thought that this did not allow to understand impact fully. In order to get a better sense of impact, a purposeful sample of CRESST journal articles and technical

reports was generated for further analyses. This sample was selected by asking four CRESST partners -- Ron Dietel, Joan Herman, Robert Linn, and Lorrie Shepard -- their opinions of the most "important" 15 or so articles from a list of CRESST journal articles and technical reports. Our final list was based largely on the list of articles generated by Robert Linn, CRESST co-director, with the addition of a few articles suggested by other key CRESST partners. The only criterion we imposed on the final list was that the articles had to have been published early enough in CRESST's funding cycle so that they would have had time to have been cited by other authors. This resulted in a sample of fourteen articles authored (only counting first author) by twelve different CRESST partners (see Appendix D).

All articles from the purposeful sample were used for a more detailed analysis in an attempt to understand how CRESST work was being used. This analysis involved reading the CRESST article so that the evaluator understood the main points of the article, then scanning the citation article to comprehend its gist, and finally, judging how the original CRESST work was used in the subsequent article. Using the list of citations generated from the first phase of our analyses we searched for and scanned the articles that cited the original CRESST work. We developed the following scale to categorize the degree of use:

1. *The article was cited as an example of general point, but was not a crucial use of the original article, in that many other articles could have been used in its place.*

2. *The article was used to substantiate a specific point in the subsequent work and was crucial to making this point.*

3. *The CRESST article serves as an important foundation for ideas developed in the subsequent work.*

Three CRESST articles/monographs (Linn, Baker, & Dunbar, 1991; Herman, Aschbacher, & Winters, 1992; Shavelson, Baxter, & Pine, 1992) were used to pilot this rubric and generate a set of benchmarks, i.e., articles that could be classified unambiguously into one of the three categories. Each article was rated by two evaluators, both of whom were Ph.D. students in education, one in educational evaluation and one in educational assessment. Any disagreements in ratings were resolved through deliberation, re-reading the citation, and comparing them to the benchmarks. A complete listing of all citations, including status and use ratings for the articles from the purposeful sample can be found in Appendix E.

The last phase of these analyses was designed as a "check" on the influence of CRESST research on the measurement community using two different approaches. First, we used the Silver Platter CD-ROM software and the ERIC system to search for the term "performance assessment," limiting our search to 1994 and the first 3 months of 1995. This search yielded 54 published journal articles. The first step of this analysis was to count the number of CRESST partners appearing as authors in this list of 54 articles. The second step was designed to see how often CRESST partners were cited in these works. This involved using the SSCI index to search obtain the reference lists for each article. For journals not indexed by SSCI, we tried to find a "hard copy" of these articles from the main library at the University of Colorado, Boulder. Using these strategies, we located 35 of the 54 articles. We searched the reference lists from these articles for the appearance of CRESST research. We

recognized that CRESST work is cited at a higher rate by other CRESST partners than by non-CRESST researchers and counted the number of citations separately for CRESST and non-CRESST authors.

The second part of this analysis involved using "First Search/Uncover" software to search for the terms "performance assessment education" for the years 1998-1995 to see if we could detect a trend in the amount of published performance assessment articles. After obtaining the search results for each year, we counted the number of articles authored by a CRESST partner (first author only). This analysis was conducted to see if there is a correlation between the number of published performance assessment articles and CRESST's years in operation and to determine if CRESST partners are representing an increasing share of journal contributions.

Results and Discussion

CRESST Publications. These analyses indicate that CRESST partners are quite productive and publish their works in fairly influential (high status) outlets (see Table 1). The two co-directors, Robert Linn and Eva Baker, were the most productive with 34 and 23 first-authored publications, respectively. Essentially, all of the books were rated "3"and chapters were given an influence rating of "2." Technical reports were excluded from this portion of the analysis. Therefore, average ratings greater than 2.0 for any author indicated that they tended to publish in the most influential journals.

While some bibliometricians dismiss publication counts because of lack of information about research quality (compared to quantity), we find the peer review process associated with the most influential journals to be a reasonable indicator of quality. Many technical reports and book

chapters are peer-reviewed but not with the emphasis on external review common for major journals. While the publication analyses are preliminary (awaiting the rest of the vita), if these patterns continue, we could conclude that most CRESST research is of high quality as determined by "blind" and often rigorous peer review.

Table 1
CRESST Assessment-related publications since 1990 and status ratings.

CRESST Partner	Total Publica-tions	Journal Articles	Books	Chapters	Technical Reports & Mono-gra	Average Status Rating
Pamela Aschbacher	14	5	0	0	9	2.0
Eva L. Baker	23	11	0	13	0	2.0
Leigh Burstein*	10	1	0	6	3	2.1
Robert Glaser	17	7	0	10	0	2.1
Edward H. Haertal	22	6	0	9	7	2.3
Joan L. Herman	21	8	1	8	4	2.1
Daniel Koretz	19	6	0	4	9	2.5
Robert Linn	34	21	1	12	0	2.4
Lorraine McDonnell	10	2	0	2	6	2.5
Robert Mislevy	23	16	0	6	1	2.4
Bengt Muthen	19	10	0	6	3	1.9
Lauren B. Resnick	26	8	0	17	1	1.9
Richard J. Shavelson	25	12	2	10	1	2.0
Lorrie Shepard	22	11	1	9	1	2.3
Mary Lee Smith*	6	3	1	1	1	2.8
Richard E. Snow	23	6	0	15	2	2.0
Noreen M. Webb	7	5	0	2		2.5
CRESST Totals	**321**	**138**	**6**	**130**	**48**	**2.2**

*Vita were unavailable for Smith and Burstein (who passed away in 1994) so their publication records were obtained from an earlier CRESST continuation proposal (1994) and an ERIC search.

Citations of CRESST research

A citation analysis performed on eighteen CRESST partners (the 17 listed in Table 1 plus Stephen Dunbar) was used to examine the influence of CRESST research (see Table 2). These eighteen CRESST partners produced 90 articles, books, book chapters, and technical reports that were cited by other researchers. These 90 articles were cited 424 times since 1990, the earliest any of these publications was released. In fact, most reports were published in 1991 or later. A substantial portion of these citations appeared in articles written by other CRESST partners (24%) or by the author of the cited work "self-citation" (17%). Nevertheless, a majority (59%) of these citations appeared in articles written by researchers not directly associated with CRESST.

To understand how often these articles were cited, once they were cited, the average number of citations/per article cited was calculated. In general, the number of citations per article is not normally distributed, rather has a strong positive skewness with the majority of articles receiving no citations and most cited articles receiving only one or two citations (Hamilton, 1990). For example, of the 81 citations for eleven articles for Robert Linn, 46 were for a single article (Linn, Baker, & Dunbar, 1991).

The work of the CRESST partners examined here suggests that this work is cited at a slightly higher than typical rate with an average of 4.7 citations per cited article. Discounting the self-citations still resulted in an average of almost four citations per article cited. Focusing on citations of articles cited at least once could lead to an overly favorable impression. Therefore, we used the information presented in Table 1 to produce a more realistic estimate of citation rates. The seventeen

CRESST partners (excluding Dunbar) produced a total of 321 articles from 1990 to 1995, 89 of which were cited at least once. These eighty-nine articles were cited 424 times resulting in an average of 4.7 citations per cited article. Dividing these 424 citations by the 321 articles written by these authors yields a rate of 1.3 citations per article written.

The last column of Table 2 contains status ratings for the citations. The same "influence/status rating" scale used earlier was used to rate the status/influence of the journal containing the citation of the CRESST work. This provides another source of information about the quality of the CRESST work. If CRESST research was only cited in "low status" journals, one could infer that the measurement community did not think it was important. Conversely, if CRESST work is consistently cited in the most important journals , one could infer that CRESST work is considered important.

As seen in Table 2, these CRESST articles were cited in influential journals with an average influence rating of 2.3 (out of a possible 3). Some of the lower ratings in this column are an artifact of the rating scale such that a more widely cited article might "lose" points because of citations in practitioner journals. This broad appeal should not be viewed negatively, but examined in the context of the other information presented in this and other tables. Nevertheless, the relatively high average rating of citation sources (2.3) is evidence that CRESST work is considered important to other researchers.

[Insert Table 2 Here]

Use and influence of CRESST research

The analyses discussed in this section can help disentangle some of the confounding impressions of how a wide appeal might lead to lower

Table 2
CRESST Citation Rates and Status Ratings

First Author	Number of Articles Cited	Number of Citations				Average # Citations Per Article Cited			Average Status Rating Per: Citation
		Total	Non-CRESST	CRESST	Self	Total	CRESST	Self	
Aschbacher	4	14	3	6	5	3.5	1.5	1.3	2.4
Baker	10	33	6	8	19	3.3	0.8	1.9	2.2
Burstein	1	2	2	0	0	2.0	0.0	0.0	2.0
Dunbar	1	9	4	4	1	9.0	4.0	1.0	2.8
Glaser	8	29	25	4	0	3.6	0.5	0.0	1.8
Haertel	5	6	4	2	0	1.2	0.4	0.0	2.8
Herman	6	16	7	6	3	2.7	1.0	0.5	2.1
Koretz	5	17	8	7	2	3.4	1.4	0.4	2.6
Linn	11	81	58	10	13	7.4	0.9	1.2	2.2
McDonnell	5	6	3	2	1	1.2	0.4	0.2	2.7
Mislevy	2	2	1	1	0	1.0	0.5	0.0	2.0
Muthén	1	6	6	0	0	6.0	0.0	0.0	2.5
Resnick*	1	24	10	13	1	24.0	13.0	1.0	2.6
Shavelson	8	52	24	10	18	6.5	1.3	2.3	2.5
Shepard	8	58	41	14	1	7.3	1.8	0.1	2.2
Smith	3	36	31	3	2	12.0	1.0	0.7	2.4
Snow	5	16	12	0	4	3.2	0.0	0.8	2.4
Webb	6	17	5	11	1	2.8	1.8	0.2	2.2
CRESST Totals	90	424	250	101	71	4.7	1.1	0.8	2.3

*Still need to do SSCI to complete Resnick. Only Resnick article reported here is Resnick & Resnick, 1992.

status/influence ratings. In this section we focused on fourteen key CRESST articles to understand how these articles are used by other measurement researchers (see Table 3). All except two of these nominated articles were cited at least six times. In total, these fourteen articles were cited 175 times, accounting for 41% of all of the citations received by the ninety articles included in Table 2. These fourteen articles received the same proportion (24%) of within-CRESST citations as the general group of CRESST articles, but they were self-cited at a slightly lower rate (11%) and cited by non-CRESST researchers at a slightly higher rate (65%) than the larger group of articles. This indicates that they were slightly more important to the larger, non-CRESST, measurement community than the ninety articles discussed above.

To compare the citation rates of these fourteen articles, the average number of citations per year was calculated. We used the number of months, rounded to the nearest quarter year, from the journal publication until the end of March, 1995 as the denominator in this equation. We used the end of the year of publication as the starting point for books. The *Educational Researcher* article by Linn, Baker, and Dunbar (1991) received more citations (46) and more citations per year (13.1) than any other articles. The articles and chapters written by Glaser (1990), Resnick & Resnick (1992), Shavelson, Baxter, and Pine (1992), and Shepard (1991 & 1993) all received twelve or more citations. The journals where these citations appeared had a high average influence rating (2.3), but this was the same average influence rating as the full set of ninety articles discussed above, indicating that CRESST articles, in general, are quite influential.

[Insert Table 3 here]

Table 3
Use and status ratings of "key" CRESST publications.

Author (s)/ Year of publication	Number of Citations			Citations/ Year*	Average Status (# rated)	Average Use Rating (# rated)
	Total	CRESST	Self			
Aschbacher (1991).	**6**	3	**1**	1.7	2.8 (5)	1.8 (5)
Baker, O'Neil, & Linn (1993).	**6**	2	**3**	4.0	2.0 (3)	1.5 (2)
Burstein, Oakes, & Guiton, (1992).	**2**	0	**0**	0.8	2.0 (3)	2.0 (2)
Dunbar, Koretz, & Hoover (1991).	**9**	4	**1**	2.6	2.8 (8)	1.5 (6)
Glaser (1990).	**14**	0	**0**	3.1	1.9 (14)	1.4 (10)
Herman, Aschbacher, & Winters (1992).	**9**	3	**1**	3.6	1.9 (8)	1.7 (6)
Linn (1993).	**7**	1	**2**	4.7	1.8 (6)	n/a
Linn, Baker, & Dunbar (1991).	**46**	4	**8**	13.1	2.3 (38)	1.8 (33)
Mislevy (1992).	**1**	1	**0**	0.4	3.0 (1)	n/a
Muthén (1991).	**6**	0	**0**	1.7	2.5 (6)	n/a
Resnick & Resnick (1992)	**24**	13	**1**	9.6	2.6 (23)	n/a
Shavelson, Baxter & Pine (1992).	**12**	3	**2**	4.8	2.3 (10)	1.9 (7)
Shepard (1991).	**20**	7	**1**	5.7	2.2 (19)	2.3 (7)
Shepard (1993).	**13**	1	**0**	8.7	1.8 (13)	n/a
CRESST Total	**175**	**42**	**20**	**62.8**	**2.3 (157)**	**1.8** (78)

The main focus of this analysis was to try to understand how these CRESST articles are being used by other researchers. While we used a three point scale to rate centrality of "use," very few citations received a "3." Most citations merited a rating of "1" or "2," so average use ratings approaching 2.0 should be considered an indicator of a "useful" article. One finding obscured by Table 3 is the difference in the way that empirically articles and conceptual or theoretical articles were used. Articles reporting new empirical findings or summaries of empirical findings tended to be cited to support a specific point and be rated "2." On the other hand, important conceptual articles, such as Linn, et al. (1991), tended to have more "3" ratings than the empirical articles, but also tended to have more "1" ratings.

In order to give a better understanding of this rating system and how these articles are used, we present examples below. The major criterion for an article to be rated "1" was that another article could easily have been use in its place. For example, it was not essential for Guskey (1994) to use Linn, et al. (1991) to make the following point:

> *Collectively, these measures are referred to as authentic assessment because they are valuable activities in themselves and involve the performance of tasks that are directly related to real-world problems (Linn et al., 1991, cited in Guskey, 1994, p. 51).*

or for Gaskins, et al. (1994) to refer to Shavelson, et al. (1992) to make this point:

> *Furthermore, assessment of students; progress must mirror these curricular emphases (Shavelson, Baxter, & Pine, 1992, cited in Gaskins, et al, 1994, p. 1041).*

For a citation to receive a "2" rating, the CRESST article had to have been critical for substantiating a specific point in the subsequent article and was critical to making this point. Direct quotations almost always received a "2" rating, but occasionally if the CRESST research was cited repeatedly throughout an article, it was often rated "2." The following examples illustrate the types of citations that were rated "2."

> Shavelson, Baxter, & Pine (1992) noted that student performance may vary greatly from one task to another, which leads to questions abut the reliability of student level scores when scores are based on relatively few performance tasks (Taylor, 1994, p. 235).

> Recently, measurement scholars (e.g., Linn, Baker, & Dunbar, 1991; Messick, 1989) have begun to include discussions of consequential validity in treatments of test characteristics (Garcia & Pearson, 1994, p. 349).

Taylor (1994) is using the Shavelson, et al. article, which was a summary of a series of empirical investigations about the generalizability of performance assessments, to substantiate a specific empirically-verifiable point. Garcia & Pearson, on the other hand, used the Linn, et al, article to lead into a discussion of consequential validity; a concept that these CRESST researchers helped make more widely known. Both citations were worthy of a "2" rating, but it is clear that they were viewed importantly by other researchers for differing reasons.

The last case demonstrates how Moss (1992) used the Linn, Baker, & Dunbar article as an important foundation to her argument. She devotes an entire page in her *Review of Educational Research* article to a discussion about the validation criteria outlined by Linn, et al, as well as including

13 direct quotations. These criteria formed the crux of the Linn, et al. article, and Moss used these validation criteria as a foundation for one strand of her main argument. This Linn, et al. article was used extensively by five other researchers, the largest number of "3's" of any article in this restricted sample. The relatively high use ratings for the fourteen articles in this analysis demonstrates that other measurement and education researchers find CRESST research important to their work.

In order to determine the boundarires of CRESST influence on this community, we conducted an ERIC search for the term "performance assessment," limited to journal publications in 1994 and 1995. There were fifty-four performance assessment articles published during this 15 month period, 10 of which were authored by CRESST partners. Having just under 20% of the articles published in this field by a single research program is an indication of the amount of influence CRESST has had on the research community. Twenty percent would be considered a critical mass according to most standards.

To discover how often CRESST work had been cited in this sample of 1994 and 1995 articles, we searched the reference lists of the 35 (of the 54) articles we could locate. CRESST research was cited 90 times in the nine (of 10) CRESST-authored publications we located and 42 times in the 26 non-CRESST authored articles. All CRESST-authored articles included references to other CRESST research, except one which was written for a practitioner audience and did not include any citations. Of the 26 non-CRESST articles, 12 did not include any post-1990 CRESST citations (though a few cited pre-1990 work of CRESST researchers), so the 42 citations from this set of articles came from 14 sources.

Among the CRESST sources, one article, Baker, O'Neil, & Linn (1994) accounted for 28 citations, far more than any other articles. If this outlier is removed, the remaining seven CRESST-authored articles had an average of approximately nine CRESST citations per article which is still more than five times the amount of citations per non-CRESST authored article. Nevertheless, CRESST was still well represented by non-CRESST authors, and it appeared that most CRESST citations were in the most influential journals of the 26. These ERIC analyses present evidence that CRESST authors contributed a critical mass of publications, and CRESST research was cited at a high rated both by non-CRESST and other CRESST researchers.

Still not feeling satisfied, we decided to conduct one more set of analyses. We noticed in the past that the Silver Platter CD-ROM ERIC software used in the analyses described above yields fewer citations than other ERIC software for the same search term. We checked this impression by using the First Search/Uncover software to document the trends in performance assessment in education publications since 1988 and the contributions of CRESST authors to performance assessment research during this time (see Table 4).

As expected, First Search yielded more articles than Silver Platter, as can be seen by comparing the 196 articles for 1994 and 1995 listed below to the 54 found for the same period using Silver Platter. Because this discrepancy seemed large, we examined the abstracts from the 47 articles listed in the 1995 search. Of these, only 23 were directly related to the intent of the search term (performance assessment education). We used this 50% correction factor for each of the years in Table 4 to present a more realistic picture of the number of articles published each year. As

can be seen from Table 4, CRESST authors had an increasing presence in the performance assessment literature since the funding period started. This analysis is only exploratory -- certainly more refinements are needed-- but both the increasing trend in number of performance assessment articles and CRESST authorship offers more evidence of CRESST's influence on the measurement field.

Table 4.
Performance Assessment in Education Articles Published Each Year Since 1988.

Year	Number of Articles Found	Corrected number	Number of CRESST authors
1988	61	30	0
1989	51	25	0
1990	70	35	0
1991	77	38	7
1992	107	53	4
1993	132	66	6
1994	149	74	11
1995 (3 months)	47	23	2

Conclusions

These bibliometric analyses offer persuasive evidence that CRESST researchers are influencing the research agenda within the larger measurement community. In isolation, any single analysis might not permit this conclusion, but taken together, the publication counts, citation analysis, use and influence ratings, and ERIC analysis all point in the same direction--to CRESST's impact on alternative assessment research.

There are several limitations to this study. The nature of the study requires the evaluator to make a number of arguable, but hopefully logical, decisions. For example, including all publications of CRESST partners assumes that all of their research is supported by CRESST when, in fact, that is not true. However, disentangling the various sources of funding to apportion credit would entail even more arbitrary decisions. While direct sources of funding might be delineated, the influence of CRESST funding on subsequent awards (i.e., funding agencies support researchers with good funding track records) would be impossible to determine. This is not an insignificant impact and distinguishing CRESST-supported research from research supported by other agencies would probably underestimate CRESST's impact.

Another limitation and cause of underestimation of CRESST's influence is the long-term nature of many CRESST research projects. By limiting our analyses to publications and citations prior to March 31, 1995, we are underestimating the true impact of CRESST. For an article to be cited by March, 1995, it would have to have been published by early 1994, meaning that the original paper would probably have to have been

prepared by mid-1993 at the latest. If these analyses were repeated in 1997 or 1998, CRESST's influence would be more apparent.

If AERA papers could be collected and their reference lists examined, this might provide a more up-to-date check on CRESST's influence, but there is no mechanism to collect AERA papers. There is a non-random submission rate to ERIC and this would yield unreliable results for this type of analysis. Therefore, the analyses reported here, though an underestimate of CRESST's influence, is the most reliable that could be accomplished at this time. The ERIC analyses offered some promise for future evaluation efforts. More work would be needed to identify search terms and sift through many "false hits." These analyses allow us to bound the field of influence.

Yet another major weakness was that we really do not have a firm basis for comparison. Influence ratings of 2.3 or citation rates of 3/year have little meaning when standing alone. Comparing these figures to the "hard" science, where most bibliometric work has been done, is not meaningful because of differing practices. Ideally, one would need to define a comparison group within educational measurement. While this is not impossible, it would be difficult because of the difference (which we could not control for) that CRESST researchers have this infrastructure and pool of $14 million. Future efforts should probably include some comparison group to help make sense of the quantitative results. Nevertheless, the multiple sources of evidence are not easily dismissed, and we can conclude that CRESST is having a definite and important impact on measurement, especially alternative assessment, research.

One final thought relates to the finding of the much higher rate of within-CRESST citations compared to citation of CRESST work by non-

CRESST researchers. There has been some attention to the evaluation of networks and programs. Averch (1990) argues that few single projects will have a major influence on the direction of science, one way or another. He suggests that administrators or policy makers can select collections of projects or "portfolios of projects" that will result in both reasonable scientific merit and produce social and economic benefits. While there are few methods to aggregate project worth to program worth, examining the connections among projects and their impact on one another in addition to their overall impact might be a useful starting point.

Similarly, science evaluators are working to develop "network indicators" to tap the extent to which the "innovative capacity depends on the quality of the relationships between its members (Ciba, 1989, p. 218). The network indicators allow assessment of the network by: (1) accurately describing the network; (2) identifying bottlenecks in communications among different stakeholders; (3) describing the action by the network to solve these difficulties; and (4) evaluating the effectiveness of actions from step 3 in solving the difficulties in step 2 (Ciba, 1989). Evaluating the effectiveness of these networks or portfolios of research is not the final answer, but again another piece of puzzle. Detailed exploration of the interaction of CRESST researcher through interviews or other qualitative methods could help evaluators identify pathways and webs of influence.

References

Averch, H. (1990). Policy uses of "evaluation of research" literature. Paper prepared under contract to the United States Congress, Office of Technology Assessment, Washington, DC.

Baker, E. L. O'Neil, H. F., Jr. & Linn, R. L. (1994). Policy and validity prospects for performance-based assessment. *Journal for the Education of the Gifted, 17*, 332-353.

Baldwin, E. & Hill, C. T. (1988). The budget process and large-scale science funding. *Congressional Research Service Review, 9, 2,* 13-16.

Brooks, H. (1978). The problem of research priorities. *Daedalus, 107,* 171-190.

Ciba Foundation (1989). General discussion II: Beyond bibliometrics. In D. Evered & S. Harnett (Eds.). *The Evaluation of Scientific Research: Ciba Foundation Conference.* Chichester, UK: John Wiley & Sons Ltd.

Garcia, G. E. & Pearson, P. D. (1994). Assessment and diversity. *Review of Research in Education, 20,* 337-391.

Guskey, T. R. (1994). What you assess may not be what you get. *Educational Leadership, 51, 6,* 51-54.

Hamilton, D. P. (1990). Publishing by -- and for? -- the numbers. *Science, 250,* 1331-1332.

Hill, C. T. (1989). How science policies are determined in the United States. In D. Evered & S. Harnett (Eds.). *The Evaluation of Scientific Research: Ciba Foundation Conference.* Chichester, UK: John Wiley & Sons Ltd.

Johnston (1990a, May). Project selection methods: International comparisons. Paper prepared under contract to the United States Congress, Office of Technology Assessment, Washington, DC.

Johnston (1990b, July). Project selection methods: International comparisons. Paper prepared under contract to the United States Congress, Office of Technology Assessment, Washington, DC.

Kruytbosch, C. E. (1989). The role and effectiveness of peer review. In D. Evered & S. Harnett (Eds.). *The Evaluation of Scientific Research: Ciba Foundation Conference.* Chichester, UK: John Wiley & Sons Ltd.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62,* 229-258.

Prewitt, K. (1982). The public and science policy. *Science, Technology, & Human Values, 7,* 13.

Sroufe, G. E. (1991). Education enterprise zones: The new National Research Centers. *Educational Researcher, 20, 4,* 24-29.

Taylor, C. (1994). Assessment for measurement or standards: The peril and promise of large-scale assessment reform. *American Educational Research Journal, 31,* 231-262.

United States Congress, Office of Technology Assessment (1991). Federally Funded Research: Decisions for a Decade. OTA-SET-490. Washington, DC: U.S. Government Printing Office.

Walsh, D. (1994, December). Common sense must tame the science genie. *The Sunday Camera.* p. D1. December 18, 1994.

Appendix A:

Journal Influence/Status Rubric

The following scale was used to rate the influence of source and citation journals. Ratings of specific journals found in this study follow the rubric.

3 = Highest status and influence journals. All AERA and NCME publications, other high status journals in education, and high status/influence journals in related fields such as psychology.

2 = Highly influential practitioner journals, but without the same status/influence on researchers as first tier research journals. This category also includes second tier research which thought of as important journals without the same status and influence as first tier journals or those with a more limited audience.

1 = Lower status research and practitioner outlets. This category included research and other journals with little status or with such a narrow audience they would not have very much influence. This category also included less influential practitioner journals than those rated '2' above. Technical reports (including most CRESST and RAND reports) are included in this category.

AERA & NCME (1st Tier) Journals
(Most prominent and highest status)

3 Educational Researcher
3 Review of Research in Education
3 Educational Evaluation and Policy Analysis
3 Review of Educational Research
3 American Educational Research Journal
3 Journal of Educational (and Behavioral) Statistics
3 Educational Measurement: Issues and Practice
3 Journal of Educational Measurement
3 Applied Measurement in Education

Other (Top Tier) Journals
(Similar status and prominence to AERA/NCME journals: in some cases, even higher rating)

3 American Psychologist
3 Educational Psychologist
3 Journal of Research in Science Teaching
3 Annual Review of Psychology
3 Educational Policy
3 American Journal of Education
3 Memory & Cognition
3 Harvard Educational Review
3 Psychological Reports
3 Psychological Review
3 Annual Review of Psychology
3 Teachers College Record
3 Teaching and Teacher Education
3 Journal of Curriculum Studies
3 Journal of Research in Mathematics Education
3 Journal of Experimental Psychology

Prominent Practitioner Journals
(Highly influential practitioner journals, but not the same status/influence on researchers as 1:

2 Phi Delta Kappan

2 Educational Leadership
2 Contemporary Educational Psychology
2 The Reading Teacher
2 Young Children
2 Educational Administration Quarterly

Second Tier Research Journals
(Important journals but without the same status and influence as 1st tier journals).

2 Educational Assessment.
2 Education and Urban Society
2 Journal of Learning Disabilities
2 Journal of Educational Research
2 Elementary School Journal
2 Monographs of the Society for Research in Child Development
2 MIS Quarterly
2 Journal of Negro Education
2 Journal of Experiential Education
? Academic Medicine
2 Journal of Experimental Education
2 Personality and Individual Differences
2 Learning and Instruction
2 Research in the Teaching of English
2 Journal of Psychology
2 Journal of Reading Behavior
2 Journal of Reading

Third Tier Research and other Journals
(Research and other journals with very little status or with such a narrow audience they would

1 Journal for the Education of the Gifted
1 Alberta Journal of Educational Research
1 Computers & Education
1 ETR&D-Educational Technology Research and Development
1 Computers in Human Behavior
1 Journal of School Health
1 Focus on Exceptional Children
1 Modern Language Journal
1 Journal of Educational Computing Research
1 Education
1 Measurement and Evaluation in Counseling and Development
1 Evaluation and the Health Professions
1 Quest
1 Journal of Psychoeducational Assessment
1 Journal of Science Education and Technology
1 Learning Disability Quarterly
1 Exceptional Children
1 School Effectiveness and School Improvement
1 International Journal of Educational Research
1 Assessment in Education
1 Technos
1 Anxiety, Stress, and Coping
1 College Board Review
1 Curriculum Inquiry
1 Chemtech

1	Journal of Creative Behavior
1	Educational Technology

Appendix B

Phase 1 (SSCI): Bibliometric Protocol

1. Conduct a Social Science Citation Index search for the following names (one at a time). When doing this search, print a list of articles where these authors have been cited. We will be limiting our study to works published after 1990, so do NOT print any citations for articles published earlier than 1990. For example, if Bob Mislevy wrote an article that referred to something written by Robert Linn prior to 1990, we would not count it in this study. Although we are concerned with assessment articles for this study, do not limit the search at this stage.

1. Pamela Aschbacher
2. Eva L. Baker
3. Leigh Burstein
4. Steve Dunbar
5. Robert Glaser
6. Edward Haertal
7. Joan L. Herman
8. Daniel Koretz
9. Robert Linn
10. Lorraine M. McDonnell
11. Robert Mislevy
12. Bengt Muthen
13. Lauren B. Resnick
14. Richard J. Shavelson
15. Lorrie Shepard
16. Mary Lee Smith
17. Richard E. Snow
18. Noreen M. Webb

Chapter 3

CRESST Influence on Test Directors' Use of Alternative Assessment

Dorothy Aguilera

Draft: Jan 10, 1996

This study is part of a multiple methods project to evaluate the impact of the Center for Research on Evaluation, Standards, and Student Testing (CRESST). CRESST's mission is to develop better ways for schools to assess student performance. This particular study focused on whether CRESST influenced test directors' decisions and thinking about using alternative assessments. Test directors of school districts and state departments of education are an important group of gatekeepers, and one would presume that they play significant roles in either facilitating or impeding the development and implementation of alternative assessment techniques. To ascertain their attitudes and opinions about alternative assessment we designed a mail survey questionnaire (Appendix XI).

Test directors were asked to provide information in four specific areas:

1) test directors' use, development, or marketing of various forms of alternative assessment, plus the time frame for using alternative assessment in their institutions;

2) test directors' ratings of the usefulness of different ways of assessing student learning, plus the usefulness of alternative assessment in comparison to traditional testing methods;

3) test directors' ratings of CRESST's influence in building awareness and use of alternative assessments;

4) test directors' ratings of usefulness of CRESST materials and resources, plus other sources used by test directors in their decision making;

5) open-ended questions soliciting test directors' opinions about CRESST and alternative assessments.

We mailed 169 surveys to test directors in public and private institutions whose names were acquired from a national association of test directors. Some of these people were not test directors at the time of the survey, but had been in the past or had worked in school district testing offices. Some worked for commercial test companies who develop and market tests. We did not differentiate current directors from others in our analyses. From the first mailing we received 54 completed questionnaires; five were returned incomplete because of job changes. A second mailing yielded 47 more surveys for a total of 101 responses, a 62.5% response rate, which we deemed sufficient for our purposes.

We organized the data into eight major categories and used a statistical software program (SPSS) to run frequency counts. For our purposes we did not think further statistical analyses worthwhile. Although there are many side questions we might have addressed with these data, the main issues we wished to answer are what test directors think about and do with alternative assessment, and what influence CRESST has had on their thinking and behavior. After examining open-ended comments, we decided against conducting content analyses and used them to illustrate various points derived from the data, again focusing on the main questions at hand. The comments are presented in full in the appendices.

Assessment Use

The initial question was designed to ascertain what forms of assessment the test directors were developing, marketing or using. Twelve options were listed for their response. including the following:

1) standardized testing
2) minimum competency
3) constructed response
4) portfolios of student's work
5) self-assessment measure
6) oral discourse
7) experiments
8) writing
9) exhibitions
10) essays
11) performance tasks
12) other forms not listed here

Space was provided for test directors to explain or describe the different forms of alternative assessment their institutions had used, marketed, or developed. The results are displayed in Table 1 (Appendix I). We organized the data according to the number of responses for each form of alternative assessment and grouped these into separate categories of most frequent use (50% or more) to least frequent (less than 50%).

Table I: Frequency of Assessment Use

Freq (n=101)	%	Forms of Assessment Most Used
94	93	standardized testing
83	82	writing
74	73	performance tasks
70	69	portfolios
56	55	constructed responses
Freq	**%**	**Forms of Assessment Least Used**
46	46	minimum competency
37	37	self-assessment measures
35	35	essays
28	28	experiments
23	23	exhibitions
19	19	oral discourse
11	11	other forms

Five forms of assessment were most often used, developed, or marketed by more than fifty percent of the test directors, including standardized testing (93%), writing assessments (82%), performance tasks (73%), portfolios (69%), and constructed responses (55%). The least used forms of assessment were oral discourse (19%), exhibitions (23%), and experiments (28%), while the moderately used included minimum competency testing (46%), self-assessment (37%), and essays (35%).

Test director comments usually referred to the assessment forms they selected from the options (Appendix I). In addition, several directors listed other forms of assessment not included in this survey but ones that they had used, such as criterion reference tests, graduation competencies, extended response answers, computer tests, physical fitness tests, and thematic performance. Other comments indicated increased development

and use of alternative assessment programs in the test directors' institutions and states, sometimes stimulated by state mandates.

Several respondents stressed the importance of using both traditional and alternative forms of assessment because both were needed to measure student competency, in their opinion. One respondent stated the focus in one district was "integration" of different assessments, a topic we infer will be important in the future based on the results of this survey. Another respondent was critical of the questions in the survey that seemingly placed more value on one assessment form over another.

> We can't replace one limited approach with another...[We] must take a balanced and comprehensive approach to an assessment system and must recognize and validate how much assessment of worth goes on daily at the classroom level (must also help to improve quality of what goes on in some classrooms).

In general, these findings show that some forms of alternative assessment were being used at a high rate, but less than standardized tests (not surprisingly). Most importantly, alternative assessment was seen as an addition to traditional assessment, not a substitute for it, and most institutions used alternative assessment only as a supplement to standardized testing. In the test directors frame of reference standardized tests occupy a place not challenged by alternative assessments, though the directors are willing to accept the latter for some purposes. In particular, they see alternative assessments as being most appropriate for classroom use. Other forms are more appropriate for district or state use, in their opinion.

Time Frames

The second question asked the length of time that the test directors' institutions had been using, developing, or marketing different assessment forms. A scale using six options ranged from less than one year to five or more years. We combined the responses into two categories, 2 years or less and more than 2 years.

Table 2: Time Using, Developing, or Marketing Assessments (n= 101)

Time Period	Freq	%
less than 2 yrs	48	48%
over 2 yrs	48	48%

Test directors were evenly split regarding the number of respondents who had less than two years or more than two years experience with alternative assessments. Use of alternative assessment is relatively recent for half the directors. In addition, some test directors' comments indicated long term use (10 to 16 yrs) of writing assessment as a popular form of alternative assessment (Appendix II). Writing assessments of various kinds are far and away the previous experience test directors have had with alternative assessment, if any.

Test Director Ratings for Usefulness of Forms of Assessment

The next questions solicited test director opinions about the general usefulness of each assessment form in assessing student learning. The "useful" responses were combined, i.e., very useful, useful, and somewhat useful were combined into a useful category to simplify the findings (Full information in Appendix III).

Table 3: Usefulness of Forms of Assessment in General

Freq (n=101)	%	Forms of Assessment
94	93	standardized testing
85	84	writing
84	83	performance tasks
81	80	portfolios
79	78	essays
73	72	constructed response
72	71	self-assessment
69	68	exhibitions
66	65	oral discourse
65	64	experiments
58	57	minimum competency

Most directors believed that there was usefulness "in general" in all eleven forms of assessment. All were considered "useful" to some degree with minimum competency listed as least useful in general (57%), somewhat surprisingly. About 20% thought minimum competency testing was "not useful," while 23% had never used it or left the item blank. Although most respondents believed that these forms were useful in general, about one-tenth had never used six of the eleven alternative assessment forms at all, i. e., oral exams (11%), experiments (10%), exhibitions (9%), constructed response (9%), and self-assessment (8%).

If one looks only at items rated "very useful," the highest category of use, then writing (47%), standardized testing (33%), and portfolios (31%) were most popular. It is somewhat surprising to find writing ranked higher than standardized testing, in light of the respondents commitment to standardized tests. None of the forms were found to be completely "not useful." The purpose of the assessment, level of use, and practicality were major considerations as to usefulness.

Table 4. Usefulness in Test Directors' Institutions

Another question asked how useful various assessment forms had been in the director's own institution, as opposed to general usefulness.

Freq (101)	%	Forms of Assessment
90	89	standardized testing
77	76	writing
67	66	performance tasks
58	57	portfolios
56	55	constructed response
50	50	essays
47	47	minimum competency
46	46	self-assessment measure
34	34	experiments
34	34	oral discourse
33	33	exhibitions

Most test directors thought that six forms of assessment had been useful in their own institutions. Standardized testing topped the list (89%), along with writing (76%), performance tasks (66%), and portfolios (57%). Many had never used some types of assessment, including oral discourse (29%), exhibitions (27%), and experiments (26%), self-assessment (22%), essays (14%), and constructed responses (12%). Even though the directors judged various forms of alternative assessment to be useful in general, their institutional environment did not facilitate their use for whatever reasons (Appendix IV).

One test director said, "The utility is determined in large part by whether the intended uses and levels of reporting are at the classroom (instructional) or large scale (accountability or certification)." Another wrote that "any 'one shot' test has severe limitations. Portfolios are the best assessment of growth over time." Some commented on the importance of combining types of assessment for different purposes: "A combination is best." "Usefulness is a function of how the test suits its

intended purpose, the ease of interpretation by intended user, and the quality." Apparently, use was restricted by a number of institutional circumstances.

Validity and reliability were also considerations. One respondent said, "Standardized (multiple-choice) accomplishes content validity and reliability. Constructed response (writing composition, and essays) tell you whether students can synthesize and explain it." Another said, "The reliability and validity along with utility for decision is very inconsistent and even troublesome with many measures, given practical constraints." Yet another, "I hold a very low opinion of alternative assessment because of the major restrictions to reliability in scoring and the validity of many tasks. If reliable scoring is available, the cost is prohibitive for urban districts with limited resources."

Another wrote, "The usefulness (of performance tasks, standardized achievement testing, portfolios of student work, and self-assessment measures) depends on the quality and appropriateness of the assessment." So test directors answered differently when asked about the usefulness of assessment forms with regard to their own institution. Partly this can be explained by general vs. specific uses. They may find certain forms useful in general (potentially) but use them in their own institutions only for specific purposes at certain levels, e. g., "Alternative assessment is best used at the classroom, not the district level."

The following table highlights the differences.

Table 5: Usefulness in General versus own Institution

In General %	In Institution (n=101) %	
93	89	standardized testing
84	76	writing
83	66	performance tasks
80	57	portfolios
73	55	constructed responses
78	50	essays
57	47	minimum competency
71	46	self-assessment measures
64	34	experiments
66	34	oral discourse
68	33	exhibitions

The differences in ratings for general usefulness and institutional usefulness ranged from 4% (standardized testing) to 35% (exhibitions). The largest differences were in exhibitions (35%), oral discourse (32%), and experiments (30%), those forms deemed less useful in general (Appendix V).

Other questions concerned the usefulness of alternative assessments when testing the performance of educationally disadvantaged students. The majority of test directors believed that these assessments were useful (85%) for that purpose. "Alternative assessment is also very useful for assessing non disadvantaged students!" The term "authentic" was challenged by some, and some viewed its use as a bias in the survey: "Why is a constructed-response item 'more authentic' than a multiple-choice item?"; and "This term [authentic] lacks a definition."

Another question asked how useful alternative assessments were in comparison to traditional testing methods (Appendix VI). The majority

believed that these forms were useful compared to standardized tests (80%), with 18% saying "very useful," 38 % "useful," and 25% "somewhat useful." Only 5% said "not useful." One said, "In comparison to traditional testing methods and for determining the actual learning of students, it depends on what is being assessed." Others wrote, "I would answer very differently depending on district or classroom level. An eclectic approach is preferred. Validity varies for each circumstance and for different students." Another said, "No form of assessment is complete, and these are not really in competition. You really need to ask about utility in the context of how assessment information will be used."

Another director said, "In determining actual learning of students it depends on what is being assessed. In comparison to traditional testing methods it depends on what is assessed?" Another was frustrated with the questions in this section and said, "Your questions demonstrate a lack of understanding of the complexity of these issues, i.e., Linn, Baker, Dunbar 1991," citing an article by CRESST researchers. Another wrote, "No form of assessment is complete and these are not really in competition. You really need to ask about utility in the context of how assessment information will be used."

In general, the respondents had a complex view of traditional and alternative assessment in which each form can serve different purposes and different levels. No assessment of any kind was seen by most as being sufficient or useful for all purposes and levels. One might infer that test director acceptance of alternative assessment depends on using traditional testing as well, and that they would resist the substitution of traditional testing with alternative forms, though none said this explicitly. Reformers who see the <u>replacement</u> of traditional testing by

alternative forms are likely to find strong resistance from this important group of testing professionals.

CRESST Impact

The next questions examined CRESST's impact on test directors' beliefs about and uses of alternative assessments (Appendix VII).

Table 6: Test Directors Ratings of CRESST Usefulness

1) How useful was CRESST in building awareness of alternative assessments for testing directors?

	Freq (n =101)	%
useful	85	84

[very (30%) useful (31%) somewhat (24%) not (4%)]

2) How useful was CRESST in getting the word out about their function?

	Freq (n =101)	%
useful	84	83

[very (22%) useful (34%) somewhat (28%) not (4%)]

3) How useful was CRESST in influencing test director's thinking about alternative assessments?

	Freq (n =101)	%
useful	77	76

[very (15%) useful (36%) somewhat (26%) not (12%)]

4) How useful was CRESST in helping test directors make decisions about testing materials?

	Freq (n =101)	%
useful	63	62

[very (8%) useful (25%) somewhat (30%) not (11%)]

When test directors were asked how useful CRESST was in building awareness of alternative assessments for testing directors, the majority (84%) thought that CRESST had been "useful" in generating awareness about these assessments. The majority (83%) of test directors also believed that CRESST was "useful" in disseminating information about their function. One respondent said, "CRESST staff have presented excellent papers and workshops at conferences." Test directors also thought that CRESST was useful in influencing their thinking about alternative assessments (76%).

From the more detailed analyses, CRESST was more successful in stimulating awareness and disseminating information and somewhat less so in influencing directors thinking or influencing decisions. This seems only natural. The directors' thinking would be influenced by many other sources of information as well, and decisions about test use are based on local factors within the directors' organization, as they themselves indicated. Even so, CRESST was still perceived as significantly useful even in these more context-specific situations. The figures for "not useful" seem surprisingly low. Most (62%) believed that CRESST was useful in helping them make decisions about testing materials.

On the other hand, several comments revealed that some did not

know about CRESST. One director wrote, "What is CRESST?" and requested the address. Another issue was the need for CRESST to develop more direct communication. One director wrote, "More direct contacts which link local efforts and concerns with knowledge, products, and resources would be helpful." Several congratulated CRESST's facility in building awareness and disseminating information. Comments about CRESST were generally very positive and stressed the high quality of CRESST work and its utility for information, clarification, and legitimation.

- CRESST was very useful in terms of potential; much of this potential has yet to be realized. The literature published by CRESST influenced my thinking about alternative assessments which undoubtedly affected my decisions.
- Materials produced by CRESST have a sound research base and are objectively presented, unlike a lot of alternative assessment resources....The credibility of CRESST personnel and the widespread dissemination of CRESST publications have been useful.
- CRESST has been a major teacher in awareness; CRESST has been the leader nationally in practical research on alternative assessment. The quality of their work is outstanding.
- CRESST has consistently been on the "cutting edge" and is especially proficient in working with practitioners about what is going on in the "real world" of schools.
- It lets me know I'm not alone out here. It's evidence I can use with the power brokers.
- ...The materials available through CRESST have been very helpful in helping us identify and think about issues.
- CRESST is considered the leader in the field of alternative

assessment.

Of course, not eveyone was happy with CRESST. There were negative comments as well, though these were a small proportion of expressed opinions.

- CRESST ...work has set us back 5 years in measuring the effectiveness of education.
- CRESST needs to get more practical by involving practitioners around the country in project discussions.
- Several years ago, Burstein moved to make CRESST a university/school partnership. It has since become the usual top-down university controlled organization.

Usefulness of CRESST Resources

Other questions asked directors to judge the usefulness of specific CRESST resources by choosing the level of usefulness for five CRESST resources (Appendix VIII):

1) technical report
2) newsletters
3) internet services
4) media products, video tapes, and database
5) document, A Practical Guide to Alternative Assessment

Table 7: Usefulness of Resources

Freq (n =101)	%	Resources Useful
79	78	Newsletters
68	67	Technical Reports
52	52	Document (book)
22	22	Media products
19	19	Internet services

The majority of test directors thought that CRESST newsletters were "useful" to them. To a lesser extent, technical reports (67%) and the book (52%) were useful. By contrast, most had not heard about CRESST Internet services or media products (Appendix VIII). (The CRESST book by Herman, Aschbacher, and Winters was widely circulated in the tens of thousands all around the country. For a fuller analysis of its impact, see the tracer studies in the next chapter of this report.)

Table 8. Contact with CRESST Professionals

Freq (N=101)	%	Contact
78	77	Professional Journals
60	59	CRESST Materials
31	31	Principal Investigators
28	28	CRESST staff
26	26	CRESST Conferences
51	51	Other sources

Most test directors (77%) had contact with CRESST through professional journals and materials (59%). About a quarter had contact

with CRESST staff (28%), principal investigators (31%), and conferences (26%). Journals seem to be the primary contact, however (Appendix IX). Comments focused on the ways CRESST had influenced ideas through workshops, journals, and newsletters. Some wrote about CRESST building awareness. One wrote, "You [CRESST] make a great effort both to pursue quality research and to disseminate information to a wide audience of practitioners. Your efforts at dissemination...is appreciated."

In general, one has the impression that most contacts with CRESST are through impersonal sources, including the usual outlets for research, and that these outlets have been successful in influencing opinion within this group. As can be seen from comments, CRESST is very highly regarded by many, though not by everyone.

- AERA presentations are the best source of information.
- We have used some of the resources identified or available through CRESST, but CRESST has only occasionally been our first source of information.
- Our materials were developed internally, and CRESST materials were useful in thinking about issues and ways to approach the task.
- CRESST has the reputation of a very high quality, professional organization. Joan, Eva, Bob Linn, etc., are highly thought of in the testing community.
- You have great researchers.
- Newsletters have helped keep a sense of balance.
- I see CRESST as a source of thoughtful technical reaction to new kinds of testing.
- Too little, too late for the money. CRESST reports/presentations validate conclusions I reached three years earlier. They never "lead"

the way. Their conferences were always during the first three weeks that schools were in session.

- An overblown concept as it replaces other methods only to be replaced by another educational sine-wave fad!

Of course, the test directors are plugged into many other information networks. Other sources of information included the Northwest Lab, North Central Lab, Far West Lab, the CCSSO Large-scale assessment conference, NCME, AERA, major test publishers, various departments of education, and so on. In a sense, it is dificult for directors to tease out precisely which of many sources of information have contributed to their thinking, but the vast majority recognize CRESST as a powerful influence.

Conclusions

In general, from this survey it appears that CRESST has had considerable influence on test directors' beliefs and use of alternative assessment. Most directors think that alternative assessment is a good thing, in general, and somewhat less so for their own institution. It is generally true about innovations that people find more restrictions and less acceptance within their home setting. They see particular impediments to change in their organization that they don't see in general. Nonetheless, CRESST influence is surprisingly strong in influencing the thinking and decision making of these key gatekeepers. Most saw CRESST as providing valuable, high quality information and attended to it closely, the reputation of the organization being an important factor.

This is not to say that the directors accept all that CRESST or anyone else says about alternative assessment. Their commitment to traditional standardized testing is strong. They see alternative

assessments as providing additional information about student performance, not as replacing traditional assessments. They see most forms of assessment as useful in the "appropriate" time and place. This is one of the most interesting findings because it indicates the way alternative assessment is likely to be used--as a supplement, not a replacement for traditional testing.

Some forms of alternative assessment are greatly favored over others. Portfolios and written exams are favored, while experiments and exhibitions are not. It may be that portfolios and the favored forms have been more heavily publicized and discussed. Or that they are easier to do, within current resouce limitations. For most directors alternative assessment is relatively recent in their experience, with the important exception of writing assessments. Their opinions may still change after more experience and reflection.

CRESST influence seems to be exercised mostly through impersonal rather than personal contacts, mostly through journals and publications. Whether this reflects something about CRESST or the test directors is difficult to say. It appears that CRESST has focused on impersonal dissemination channels, but it may be also that test directors prefer journals and formal publications as sources of information. The long history of attempted educational innovation suggests that personal contacts are necessary to successful implementation, especially the more complex and difficult the innovation. Although the test directors are critical to acceptance and implementation of alternative assessments, personal contact may be more necessary at the teacher or classroom level, the level of those who have to do it. Our assessment of teacher acceptance, a later chapter in this report, suggests that doing alternative

assessment in the classroom is not easy. One consideration for CRESST might be to establish opportunities for more personal contacts the closer to the classroom level the activity.

In summary, to answer the major questions of this survey, test director acceptance and use of alternative assessment seem surprisingly far along, given the caveats about using alternative forms along with traditional assessment, and CRESST has been significantly and measurably influential in this acceptance and use of alternative assessment by this significant gatekeeper group. CRESST influence is based in part on the perceived high quality of information it provides and the reputation of its researchers. The credibility of CRESST R&D not only helps persuade test directors, but also helps them legitimate alternative assessment to their constituencies. The extent to which alternative assessments will actually penetrate classrooms must await further study, but acceptance by these test directors is certainly a necessary step.

Chapter 4

Tracer Studies of CRESST Products

Linda Rastelli

Abstract. Two tracer studies were conducted, using two products designed by CRESST--a book and a model for developing performance assessments. The impact of the products was explored by interviewing a sample of product users supplied by the center. The book was characterized by our informants as a very widely known and respected resource in the field of alternative assessment. The model was less well known and used, but the ideas it was based on, detailed in the book, did have influence beyond the model's immediate users.

Method

Two tracer studies were conducted to present a picture of how two CRESST products, nominated by the center as outstanding exemplars of successful products of the center's research and development efforts, have impacted their users. Organizations were the unit of analysis for a "content assessment" model designed by CRESST researchers, and individual educators were the unit of analysis for A Practical Guide to Alternative Assessment, a book about performance assessment.

The tracer studies were conducted mainly through telephone and electronic mail interviews. A few were conducted by facsimile, and one was done in person. Initial informants for the interviews, after CRESST researchers were interviewed about the development of the products, consisted of a list of major users for each of the products, supplied by the center.

The samples used in these studies were not intended to characterize representative samples of the population of alternative assessment professionals, the assessment community, or educators in general; thus very little numerical information has been used to analyze the data. To attempt to quantify our data would not be meaningful, though, because our purpose was not to conduct a quantitative survey of the products.

Instead, we aimed for depth rather than breadth of responses in an attempt to explore widely the range of use for each product and acquire insights into reason for its use, and to illuminate the path taken by the products to get to their users. An assumption was made at the outset that our respondents were to some degree satisfied with the products, or they would not be using them extensively in the first place, although we did collect information on potential problems and suggestions for improvement of both products.

Most of the data were qualitative except for a question[1] about how many recommendations an interviewee had made. After the data were collected, they were analyzed by compiling the responses into categories in order to generalize, albeit loosely, about usage and opinions about the products.

The Practical Guide

A Practical Guide to Alternative Assessment was written for an audience of "preservice and practicing teachers, school administrators, and district- and state-level practitioners who are interested in developing new kinds of assessments."[2] The book is a short (121-pages)

[1] See interview protocol, appendix, for the specific questions asked.
[2] Joan L. Herman, Pamela R. Aschbacher, and Lynn Winters, (1992), A Practical Guide to Alternative Assessment, Association for Superivision and Curriculum Development: 2.

handbook that explains how to select and score performance tasks, how to link assessment and instruction, and other related aspects of implementing authentic assessment. It relies on the center's content assessment model, a process model that links curriculum, learning and instruction. Its authors are Joan Herman, associate director of CRESST; Pam Aschbacher, project director at CRESST; and Lynn Winters, assessment director of the Galef Institute in Los Angeles.

For the Guide, the center supplied data concerning requests to use copyrighted materials, as well as names of people who CRESST researchers were aware were using the book extensively. We also sent an electronic mail request to a list of the American Educational Research Association Division D membership and asked for persons familiar with the book to participate in short interviews. Our informants were asked to provide brief comments about their usage, opinions, how they heard of and potential problems with the book. (See Appendix for interview protocol.) They were also asked to approximate how many others they had referred the book to and in what context, and to provide names, if possible.

When necessary, follow-up interviews were attempted. We then contacted this second group of people, when possible, and continued this process until we felt we had interviewed enough people to obtain a good idea of how the book was being used. At this point, 21 interviews had been completed, the leads had begun to fade, and we believed we were duplicating data. To obtain data about (K-12) classroom teachers, who were not part of CRESST's lists, a list of teacher training workshop participants was obtained from public school officials in one state, of whom five were called at random and asked to participate. Two in this

group participated by faxing written responses to the questions; one agreed to a telephone interview.

Our final sample for the guide included a total of 24 individuals. All but one were familiar with the book's contents. The occupations of the sample were as follows: Nine university professors, two full-time educational researchers, and three classroom (K-12) public school teachers. The rest (10) were educational professionals employed by public schools, four at the state level and six at the district or local level.

The Content Assessment Model

The content assessment model is a performance assessment model designed by a team of researchers led by Eva Baker at CRESST for constructing performance assessments at multiple grade levels and content areas. This framework, based on an analysis of different learning styles, was created by the center to help districts and states implement large-scale assessment systems without having to create each assessment separately.

The model is referenced in the Guide, which, unlike the model, was aimed primarily at classroom teachers, according to Baker. The model, which has been refined several times since its original inception, was developed for a "broader audience" of professionals working toward instructional improvement and system accountability, she said. This includes commercial publishers and state and district assessment professionals, as well as classroom teachers.

Information about major users of the model was also provided by CRESST staff. The institutions discussed in the study were selected for having done work with the content assessment model. After making

initial contacts with these institutions through names provided by CRESST, we broadened our inquiry by talking to other people who had been referred by our first contacts. Respondents agreed to talk openly with us in return for a promise of confidentiality that their names would not be used in the study. Thus when writing about the interviews done for the model, we have usually not identified the institution or name of the informant.

In several cases it was necessary to depart from the interview protocol questions in interviewing individuals who did not have specific knowledge of the model itself, but only of different performance assessments that had been developed using the model. In these situations, informants were questioned about their experience with the specific assessments they had worked with. There were also several individuals who had had extensive contact with and guidance from CRESST researchers and materials, but did not actually use the model to develop assessments. These interviewees were helpful in demonstrating the center's impact on thinking and practices, although they were unable to evaluate the model. Additionally, there was another category of interviewee who had contracted with CRESST to build performance assessments, but did not have enough experience with the center to offer an opinion.

Our total sample of interviewees for the content assessment model consisted of 17 individuals, of whom only eight had firsthand experience with the model and could complete the interview protocol.

Results

A Practical Guide to Alternative Assessment: An Unanticipated Best Seller

At the beginning of the decade, "alternative" or "authentic" assessment was rapidly becoming a topic of interest to many educators. New findings about the problems of standardized testing had been disseminated, and stories featuring alternative assessment were appearing with some regularity in the mainstream media. Joan Herman, associate director of CRESST, and her colleagues, Pamela Aschbacher and Lynn Winters, knew there was a need for practical, in addition to theoretical, information on the subject.

"A lot of claims were being made for alternative assessment, but practitioners were trying to do it without having knowledge of how to do it," Herman said, "We'd accumulated a beginning expertise at the center." When they wrote A Practical Guide to Alternative Assessment , CRESST researchers had no idea that it would turn out to be ASCD's[3] second highest selling title in 1992.

Today, a small public college in West Virginia is being aided by the book in writing outcomes assessment plans for all its departments. In southern California, a university professor is using the book to teach educational measurement technologies. Teacher training workshops in Illinois rely on the book, and its school improvement planning process, now underway throughout the state, is being guided by the book's ideas. The guide is widely quoted in the field of alternative assessment, according to a school official interviewed for this study.

The book began its journey by being marketed to the ASCD membership, as well as non-members, through its regular catalogs and fliers, according to Ron Brandt at ASCD. Sales have been very high, he said, with 43,650 copies of the book sold in its first two and a half years

[3] Association for Supervision and Curriculum Development, the book's publisher.

on the market. Another 90,000 copies were distributed to ASCD's comprehensive membership when the book was published in 1992. Many of these copies were distributed in Los Angeles and Hawaii for teacher training workshops. The checklists included in the book , as well as selected chapters, are frequently requested for reprinting.

Usage patterns

Many educators who are involved with authentic assessments at many levels are using the guide extensively with a wide range of students and curricula. Our informants included education professors, in-service teacher trainers, public school administrators, test directors, educational researchers, public school teachers, and textbook authors. Many of them wear several of these hats.

Many respondents said they used the book for in-service teacher training in K-12 public school systems, although several used it for college level courses and as background information in texts they were writing. Although most courses were for pre-service teachers, one was a developmental psychology course and another was a measurement course for doctoral level students. Other uses included increasing administrator knowledge, informing school improvement plans, as primary or supplemental texts in teacher training workshops, and setting standards and assessments at the state level.

A popular chapter of the book, "Insuring Reliable Scoring," was reprinted by Educational Testing Service in Princeton, N.J. in a self-published workbook, Performance Assessment Sampler, used by roughly 1,000 state education employees, state testing directors, and educational researchers.

The Illinois State Board of Education bought 6,000 copies for use in its regional training centers. An interviewee said that 23,000 copies were distributed within the Chicago public schools alone. Illinois public school administrators found the book "quite helpful" in the state's school improvement planning process and use the book in training workshops given every six weeks for school staff. This individual recommends the book for workshops and uses its checklists for overheads and presentations. All of the state directors and school improvement people have received copies of the book, and it is referenced in the state's publications. It has served as "a perfect fit" for the assessment component of the state's school improvement plan, said a Chicago official, because it helps teachers focus on the assessment of outcomes as something students can do or know, rather than as activities.

A California educator uses the guide often as a reference. Mainly, the book was helpful in devising recommendations for assessments for a program of classroom and service learning activities in an urban school system. A Florida educator uses it to train teachers who are inservice, has cited it in her work, and has recommended it to professionals inside and outside her organization. A midwestern public school official used it for developing training modules for teachers and quoted heavily from it in a book she wrote on the subject of performance assessment.

An education professor on the West Coast said the book was "very, very useful" for his class of doctoral candidates, who "enjoyed it immensely." He used the book in conjunction with a James Popham text in order to complement what he calls the "classic assessment approach" with the book's more recent ideas. He views performance assessment as

"an emerging technology" that is not a substitute for traditional techniques, but represents a promising new approach to assessment.

An educator employed by an agency that works with 40 public school districts said these districts had been "at sea" trying to develop performance assessments without enough guidance. "When confronted with performance assessment, people get insecure that they're not getting the right kind of data," she explained. The book has helped in changing attitudes and perceptions, helping link assessment to curriculum, and developing the skills of teachers and administrators, she said.

"A great book "

Many reasons were offered for the guide's acceptance. When respondents were asked to name the most helpful aspect of the guide, typical responses named its accessible writing style, practicality, charts and examples, and discussions of reliability and validity. The guide was commended for "living up to its title" by giving practical suggestions for selecting and scoring reliable assessments.

"It provides a theoretical foundation without going too far away from the practical for teachers," said an interviewee, who praised the writing style for being clear and easy to understand. "The concept of validity is nicely summarized and it has little educational jargon, unlike other educational publications," she added.

Several informants noted that the book links assessment to instruction in a way that other materials do not. "It does not make the common mistake of emphasizing instruction over assessment, but integrates the two," was one comment. Another informant, a

schoolteacher, said that one of the most helpful aspects of the book was in showing "how assessment is part of the process, not an add-on."

"It's a great book," another informant said. "It not only covers all of the major issues--validity, reliability, bias and construction of performance-based assessment. It talks about these very deep issues in ordinary language that helps teachers to understand." From a classroom teacher: "It was written so I could ask people to read a specific chapter and it made sense without reading preceding chapters."

One informant called it "the most widely quoted writing in the area of performance assessment,"[4] because of its position as "the first place, and for a long time, the only place to find very good technical information on constructing a performance assessment." "There is a big need for good stuff on performance assessment because there is a lot of bad stuff out there," another interviewee said.

Other favorable comments: "provides a good overview of the subject," "motivational," "outcome, not task focused," "consistent with the measurement literature yet extends its application," "compact," and "good for beginners."

Path of Influence

Another question explored by this study was, how did the book reach its audience? In our research, we found the book in many cases influenced administrators and teacher educators first, and then frequently its ideas reached the classroom through teacher training programs. When public schools were involved with performance assessment, it was often a path

[4] This is not to be confused with academic citations. Our informant was referring to informal spoken comments, as well as written quotations.

that started with the state's department of education, and reached the classroom through district programs to implement performance assessment through school improvement plans or curriculum frameworks.

Professional conferences and workshops given by CRESST on alternative assessment also played a role in publicizing the book, as well as informal networking at other educational conferences. Several of our respondents who were teacher educators included the book on reading lists for courses, bibliographies referenced in their own works, or incorporated the book's ideas in their own texts.

Of the informants who remembered how they had learned of the book, most read it because they were ASCD[5] members and had been mailed a copy. But many of them added that they had been seeking better information on alternative assessment prior to receiving the book. A few respondents had sought out the book actively by asking colleagues at conferences or elsewhere for information about authentic assessment or had responded to advertising about the book. The classroom teachers in our sample mainly learned of the book through teacher training courses or from state consultants.

When asked to estimate how many people they had recommended the book to, either personally or by including it in a bibliography or reading list for a class, the most frequent response was between 25 and 100 persons. Five interviewees said they'd recommended it to more than 1,000 people, and seven did not answer the question. One replied, "every time I've recommended it, they've already known about it."

---

[5] According to information provided by ASCD about their membership, in 1992, when A Practical Guide was published, there were 90,000 "comprehensive members," who received complimentary copies of the book, of whom about 50,000 were school principals, 20,000 central office administrators, 10,000 higher education, 5,000 teachers, and 5,000 other--state department, regional service agencies, board of education members, consultants, etc.

Weaknesses and Suggested Improvements

Suggestions for improvement were varied, but it is noteworthy that nearly 50% of the sample did not think the guide could be improved at all, and when asked to describe any problems they had experienced with the book, only one respondent listed any--that some of the terminology in the book was confusing when used interchangeably.

Because it is described as a basic beginning, and good for its intended purpose, its limitations are that it lacks depth and sophistication. But because its size and simplicity were seen as major assets, to expand the book (as some suggested) may threaten these strengths. One respondent, explaining that the book's strength was its brevity, suggested explicitly that the book *not* be expanded.

The most common suggestion was to include more case studies, more concrete examples, particularly first-hand information from alternative assessment practitioners. One informant suggested that the book "go further" in elaborating examples of student work scored according to criteria, and asked for a videotaped version of its ideas. Some respondents said the book could be "updated," and another asked for the results of "assessment projects." Other suggestions: an annotated bibliography, the guide's information applied to other countries and at the college level, student samples "to pull out and score," portfolio information, and a rationale for why alternative assessment is needed.

Content Assessment Model

A Blueprint for Linking Instruction, Learning and Assessments

The center's "content assessment model" is a cognitive psychology-based model that CRESST researchers developed to solve what they judged to be common problems with performance assessments--a lack of "robustness" and a lack of "generalizability" across the curriculum, according to project director Eva Baker.

Baker believes that many educators approaching performance assessment routinely start with a good idea of an interesting task, and then look for ways to score the task, without thinking closely about what "cognitive demands" are actually involved in performing the task. This practice leads to a gap between students' "concepts or cognitions" and the task specifications, so that many performance assessments do not necessarily take into account the type of learning taking place during instruction and how it relates to cognitive demands. CRESST had done the empirical work to understand this crucial "intermediary step" to get from content standards to creating assessments, she explained.

Additionally, assessments created in this way, even within the same subject area, would not be robust, in that they would not perform the same way with each other, and would be "insensitive to varying content emphasis and epistemological differences among... experts and teachers."[6]

Another problem the center discovered was that performance assessments were being "handcrafted one at a time," with little generalizability across curriculum areas. The models offer a way of "regenerating lots of assessments" so that a district or state can "focus across subject matter" without having to reinvent the wheel, Baker said.

3Learning Based Assessments of History Understanding

CRESST researchers drew on recent research about subject matter expertise and constructivist principles to delineate five main types of learning that underlie the content assessment model. These learning types--applicable across subject areas--are: content knowledge, problem solving, communication, teamwork or collaborative work, and metacognition and work habits.

Assessment Projects

There are several large-scale assessments on which CRESST researchers have collaborated, using the content assessment model as a framework for developing performance assessments. The center also sponsors workshops and conferences on designing performance assessment with the model for educators and other assessment practitioners.

In Hawaii, the state department of education worked closely with the center to develop and implement statewide assessments based on the model. A two-year pilot in the subject of history was conducted in 1994 and 1995, and language arts was added in the second year. The state is now proposing to implement the CRESST assessments as part of their statewide assessment program which will also include norm-referenced standardized testing. This program is CRESST's largest endeavor with the model.

New American Development Schools (NASDC) funds a project known as the Los Angeles Learning Centers (LALC), for which CRESST was chosen to design and administer assessments. These schools, supported by a business consortium, are designated as "break the mold" schools. CRESST's design and implementation of English and science assessments

at the tenth and eleventh grade levels in two of these schools led to a $900,000 contract with the Los Angeles Unified School District, signed in December 1995. This effort, expected to last three years, will "draw heavily" on the content assessment model in developing standards-based assessments in four content areas for what is the second largest district in the country with approximately 650,000 students.

The Washington [State] Commission on Student Learning brought in CRESST researchers for consulting on performance assessment after the state passed a law in 1993 mandating content standards and statewide traditional and alternative assessments. Several districts in the state, including the Vancouver School District, had the center's assistance in crafting assessments. Vancouver schools piloted model-based assessments for the arts, math, and social studies that CRESST researchers had helped design. This served as a "springboard" for other performance assessments designed by the district, according to a research director for the district.

The Department of Education in Missouri is in the beginnning stages of building assessments that will combine traditional test items with performance assessments, with the indirect guidance of the newly-created Center for Learning, Evaluation, and Assessment Research at the University of Missouri at Columbia (UMC). This center is linked to CRESST through its work with former CRESST researcher and performance assessment expert David Niemi. Although the state plans to employ a test publisher to create the actual assessments, and UMC has no formal influence on the assessments, it created a technical advisory committee including Niemi and others to advise the state on how to proceed.

CRESST also has conducted workshops for and piloted assessment projects with Department of Defense dependent schools (DODDS) in Germany, through a long-standing relationship with DODD's Computer Assisted Education at Technology Insertion (CAETI). Some of these projects are ongoing, including a recent one to utilize student assessments in a computer-based environment in Germany.

Project Benefits

Several of these projects are in the beginning stages and cannot be evaluated; many potential informants declined to participate in the study for this reason. What follows is a description of the comments elicited during our research. However, given the small number of individuals having enough first-hand experience with the model or model-based assessments to come to conclusions about its usefulness, this evaluation cannot make a definitive statement regarding its influence, unlike the Practical Guide.

Our informants who had worked with the model praised the assessments, and even where CRESST-designed assessments were no longer being used, most said that the center had had a strong impact on their thinking and practices about assessment. The assessments "hit all levels of students, accomodating different learning styles," "were carefully constructed and more likely [than other PAs] to provide better information of what students can do, including changes over time," "[were] a good exemplar of using primary source documents to engage kids in complex thinking and different perpectives," and "it worked."

Among our respondents, there appeared to be a conceptual understanding and appreciation for the why alternative assessment was

being used. The tests "reinforced the need for going beyond a multiple choice format to get at understanding," an educator said. Because the model is based in "constructivist learning theory," it "helped its acceptance by educators," an informant noted.

An aspect of the model that was mentioned twice as very useful was the idea of using a "cut paper" to exemplify the upper or lower limit of a score range, rather than a "model" paper that would exemplify a midrange score. "This made a lot of sense to us," said an informant. Another educator from a different institution said that this idea had helped them to fully implement their rating system by enabling them "to distinguish a 3 from a 4."

Positive comments about the center itself included references to its expertise, national connections, and ability to offer practical, disinterested advice about implementation of often controversial plans. One informant said, "They're very respected and they don't have a profit agenda. They helped me to sort out the real stuff from the bull [about assessment]." Other comments: "They were valued for their expertise and experience--for having fought the battle before," "We were able to draw on their broad-based experience," "They don't run a number on you," "We would use them again," and "Helped me to get into the discussion about assessment that I had not been privy to until this."

Nearly all of our interviewees had recommended the center's work to other educators in other states or districts. The most frequent response to a question about number of recommendations was between five and ten people outside the informant's institution.

Problems and Suggested Improvements

Although the scope of this tracer study is so limited that we are reluctant to make generalizations, there were indications that practical problems encountered during the implementation phases of assessment systems in a few cases caused doubts about the viability of alternative assessment. The following concerns were voiced:

- *The assessments] were too advanced.*
- *Training [teachers] to score [assessments] was very time consuming.*
- *They may have solved the academic problems but not the implementation problems.*
- *It's a bit highbrow. It's good for state, but not individual data. Too much of a literacy assessment.*
- *Difficult to assess over multiple days. Very difficult for teachers to score.*
- *In reality it's [viable PA] a long time coming.*
- *Teacher involvement needs to be emphasized more.*
- *A tremendous burden for schools. Needs refinement. Is there a middle ground?*
- *[State] politics kept CRESST from being as fully involved as they could be.*

Many of these concerns, of course, were not within the control of CRESST, but are difficulties caused by political factors, or problems with performance assessment in general. Other problems noted by respondents were logistical problems in communicating with the center or in "follow-

through" on projects. These statements were qualified by respondents, however, by saying that they did not hold the center responsible and these difficulties did not influence their desire to continue working with or work again with the center.

Discussion

What the tracer studies demonstrated was that the center's products, particularly the Guide, have served to provide highly useful, concrete information to educators seeking to implement alternative assessment. In the majority of cases, the users of both products were already convinced of the benefits of alternative assessment before coming into contact with the products.

The most common situation for our respondents appeared to be that the products came into use for assessment projects that were already underway. For example, in Washington, members of the Commission on Student Learning who had been asked to begin developing and implementing performance assessments required by a new law contacted CRESST (whose reputation was known to them) to help them determine how best to do so.

In this case, among others, the center's products met a defined need that had already been identified by educators, who knew of CRESST's work in alternative assessment and went to them for this reason. In many cases with both the book and the model, informants said they had been supportive of alternative assessment but lacked the tools for setting up the system.

We did not see evidence of these products changing the oppositional views of informants; rather the products served to clarify and reinforce the users' rationales for using alternative assessments. With the model in particular, new concerns were raised about the viability of

implementing such plans. These concerns however, were directed not toward the theoretical benefits of performance assessment, but were rather issues with implementation. This is because these users had turned to CRESST at the implementation stage.

We see CRESST's role in the assessment community as extending beyond merely supplying products, but in meeting the needs of educators for practical guidance and support. The center seems to have taken on this role very successfully. CRESST's expertise was highly valued, but perhaps most important was that the products came in at the crucial and grueling implementation phase of the policy cycle, when the best laid plans often founder on the day-to-day difficulties of change. Given this challenge, it is perhaps inevitable that the model would run into criticism, however, as previously stated, given its limited usage at this time, it is not possible to draw definitive conclusions about its value. The book, however, can be viewed as a clear success. Having drawn a highly enthusiastic following, its influence appears solid. Also, because the book draws on the ideas of the model, the model can be considered to have an indirect influence in this way. Whether the model itself is viable in the large scale assessment systems for which it was designed is yet to be determined; the center's recently finalized contract with the city of Los Angeles should provide further evidence of this.

Appendix

Interview Protocols

Practical Guide interview protocol:

1. What aspect of the book have you found the most helpful?

2. Please describe and give examples of how it has influenced any projects.

3. How did you learn of the book?

4. How many people have you recommended it to, inside or outside of your institution?

5. Please describe any problems you've encountered, or explain how you think the book could be improved.

Content Assessment Model interview protocol:

1. What aspect of the model/assessments have you found the most helpful?

2. Please describe how CRESST has influenced any assessment projects.

3. How did you learn of the model/center?

4. How many people have you recommended CRESST's work/model to, outside of your institution?

5. Please describe any problems you've encountered, or explain how you think the model/assessments could be improved.

Chapter 5

CRESST Influence on Teachers

Tim Weston

Abstract. The University of Colorado CRESST project was an effort to research and promote the use of performance-based assessment in three elementary schools. The CRESST project is briefly summarized with special attention to the changes in assessment, instructional practices, and beliefs arising from the CRESST intervention, along with recommendations by researchers about the obstacles and lessons learned during the course of the CRESST project. Three teachers who participated were interviewed to learn their views of the benefits of working with performance-based assessment.

Teachers were also asked about the lessons they learned and difficulties they experienced. They reported the continued use of different forms of performance assessment and an overall positive evaluation. Several reasons were cited for continued use of performance assessment, including: 1) better ability to diagnose and understand student achievement and thinking, 2) more student self involvement in their own assessment, and 3) easier and more detailed communication with parents about student achievement. Teachers reported using a variety of performance-based assessments in their classrooms, along with the continued use of some traditional assessments.

CRESST investigator Hilda Borko recommeded 1) prior agreements should be made before the outset of research with participating teachers calling for overt attention to beliefs about assessment and instruction, 2) a clearer idea of the time and resource expenditures needed to successfully complete the project should be communicated to teachers at

the outset of the project, 3) more released time for teachers to complete project tasks should be scheduled, 4) a longer time (more than one year) should be spent by teachers in professional development projects, and 5) classroom observation and mentoring should be part of any professional development research effort.

The CRESST Project

A question remains as to how teachers will react to alternative assessment. The purpose of this paper is to describe the lessons learned by researchers and teachers during a pilot research project developing and implementing alternative assessment in classrooms, and to describe the impact the intervention had upon teachers' own assessment and teaching practice, from the teachers' point of view .

The University of Colorado CRESST project (conducted from 1992 to 1995) was an effort to research and promote the use of performance-based assessment in the classrooms of three schools. Hours of discussion, interviews, analysis, reflection, and writing went into producing CRESST publications, presentations, and technical reports about the intervention. Four University of Colorado faculty CRESST partners -- Laurie Shepard, Hilda Borko, Elfrieda Hiebert, and Bobbie Flexer -- as well as several graduate research assistants worked on the project. Fourteen third-grade teachers participated. These teachers devoted a great amount of time and effort discussing assessment in workshops and developing and implementing new forms of assessment in their classrooms.

For two years, CRESST researchers worked with classroom teachers to develop performance-based assessments that fit the teachers' curricular goals in math and reading: these curriculum frameworks reflected national standards (i.e., NCTM standards) that call for higher-

order thinking and problem solving in these subject areas. The impetus was a dissatisfaction with large-scale, multiple-choice testing and its impact on classroom instruction. Traditional tests have been criticized by CRESST researchers (and others) for directing classroom instruction toward "lower levels of thinking" (e.g., rote learning, memorization) to match the content of high-stakes multiple-choice tests (Flexer, et. al,1994). Part of the CRESST intervention was to free classroom teachers from their yearly multiple-choice test for the duration of the research, and let teachers develop meaningful performance-based assessment outside a high-stakes environment.

To understand the goals of CRESST researchers it is necessary to review two different theories about the implementation of performance-based assessments, so-called "top-down" and "bottom-up" models of assessment reform. The top-down approach is directed towards the use of state or national assessments that call for higher-order reasoning and problem solving. In this model, teachers prepare for the test by altering their instruction to emphasize the skills found on the new, more authentic test. The bottom-up approach, favored by the CRESST team, helps teachers "change their assessment program in ways that comply with the Standards ... and change their instruction to align it with their assessment" (Flexer, et. al, p.2). Use of performance-based assessment (and resulting changes in instructional practices) would lead to better achievement by students on any mode (multiple-choice or alternative assessment) of ability test. While teachers who develop their own assessments may still prepare their students for high-stakes multiple-choice tests, preparation would consist of only "test-wise" activities and

avoid the negative instructional effects associated with extensive (and distorted) test preparation.

The CRESST intervention was both professional development and a way of collecting data. The effort is described in detail by numorous CRESST publications and technical reports (see 1995 CRESST Product Catalog for information). In the CRESST paper entitled, "How 'Messing About' with Performance Assessment in Mathematics Affects What Happens in Classrooms," (Flexer, et. al, 1994), the intended effects of the research were "...to help teachers change their assessment practices" and to "expand classroom assessment repertoires, e.g., by helping [teachers] learn to design and select activities, develop scoring rubrics, and make informal assessments 'count'." There is a similar statement in another CRESST report:

> Our initial intention was to facilitate changes in the teachers' assessment practices by helping them to think about their instructional goals and the relationships among goals, instruction, and assessment: to develop or select assessment tasks appropriate to their goals; and to articulate scoring criteria for the assessment tasks. We expected that each team of teachers ...would design or select a shared set of assessments that reflected key goals of the school and district in mathematics and literacy, and that individual teachers would adapt assessments to their particular classroom contexts" (Borko et. al, 1995).

The intervention proposed by the researchers was an extensive program of staff development meant to bring about the proposed changes

in practice. The defining features of the intervention were clearly stated by the researchers.

> The intervention or staff development included several full- or half-day in-service workshops attended by teachers from all three schools, the biweekly workshops within schools, project "assignments" that each teacher did with her class between workshops, demonstration lessons in two schools and consultation on making observations in the third. Three interviews that were part of data collection ...are also part of the intervention because they gave teachers a chance to reflect formally on their beliefs and practices (Flexer, et. al, 1994).

The Shepard et. al paper (1994) contains a similar description of the research intervention:

> The intention of the project was not to introduce an already-developed curriculum and assessment package. Rather, we proposed to work with teachers to help them develop (or select) performance assessments congruent with their own instructional goals...We met with teachers for planning meetings in Spring 1992-93 school year, alternating between reading and mathematics so that subject-matter specialists could rotate among schools (Shepard, et al., 1994, p.7).

The beliefs and instructional practices of teachers were also examined by the CRESST researchers. They state: "We did not intend to confront directly teachers' beliefs but expected beliefs would shift through work on assessment practices and, as it turned out, on instruction practices" (Flexer et. al, 1994, p. 3). The researchers posited a two-way

causal relationship between belief and practice. While the researchers hoped (in some cases) that teacher beliefs would change, the researchers did not plan an intervention to directly change teachers beliefs. This restriction was codified in prior agreements made by the researchers with the school district and the teachers.

The CRESST technical reports describe the impact of the research intervention on participating teachers. Like many social science endeavors, the "treatment effect" of the CRESST intervention was neither uniform nor simple. Concrete changes in assessment and instructional practices were recorded, but these changes were superficial unless accompanied by an understanding of the beliefs, thinking, and philosophies that support the use of different types of assessment and instruction. For some of the teachers, beliefs about assessment and instruction changed, though, in general, changes in assessment practice reflected the preexisting beliefs held by the teachers, as attested to in the Borko et al. (1995) technical report:

> In general, we found a pattern of changes consistent with preexisting beliefs at all three schools. When teachers' initial beliefs were compatible with the reform agenda of the CU team and the wider mathematics education community, they implemented new practices suggested by the CU team that were related to these beliefs fairly easily and quickly (e.g., requiring explanations to accompany problem solutions). On the other hand, when teachers' beliefs were not compatible with this reform agenda, they either ignored ideas (the continued use of chapter tests by several of the teachers at Pine) or inappropriately assimilated them into existing

practices (e.g. scoring rubrics used by 2 teachers at Spruce that included spelling and punctuation...)

Some assessment and instructional practices changed, while others did not. The effects of the intervention varied more than the researchers expected: each school implemented different types of assessment to varying degrees, and individual teachers at the same schools became interested in different types of assessment and implemented these assessments with varying levels of conceptual understanding.

A few generalizations can be made about changes in assessment and instructional practices. In math, participating teachers working together with researchers developed their own performance-based assessments, and rubrics to score these assessments. Teachers also designed their own methods of observational record-keeping. Assessments in math stressed problem solving and open-ended explanations of student reasoning over traditional "one right answer" calculation worksheets and math tests. Rubrics (or scoring guides) to assess the adequecy of student explanations and skills were successfully developed by many teachers, and implemented in classrooms.

The use of observational record-keeping to track student progress was less successful than other forms of assessment because many teachers found these recording systems to be unwieldy, awkward, and time consuming. In addition to the use of alternative forms of assessment, researchers also reported that teachers made a shift in their day to day interaction with students. According to the researchers, teachers "began to ask different types of questions-- questions that encouraged students to explore and articulate alternative problem solving strategies rather than directing them towards finding the correct answer"

(Borko, et al., 1995). Some teachers also made shifts in instruction, as "problem-solving and explanation [became] much more central components of their mathematics programs" because of the intervention.

In reading, researchers met with teachers at the outset and identified goals for reading instruction, along with assessment and instructional tools for achieving these goals (Hiebert & Davinroy, 1993). Four types of assessment/instruction were identified. Summaries called for the ability to understand and synthesize text; students wrote summaries of articles, stories, or other texts and scored these efforts with a rubric. Running records display a students' progress reading aloud where errors are tracked and compared over the school year and instruction is tailored to fit each child's needs. Literature logs, a student's record of their own reading, are a means to self-reflection and self-evaluation as students read literature. Finally, annotations are anecdotal notes that teachers make about student's progress. Researchers say that the participating teachers tried each form of assessment, though the implementation was often "by the numbers" and somewhat formulaic. Each type of assessment was used by teachers with varying degrees of success. (Personal communication, Davinroy, 1996).

One goal of the reading researchers (and in math as well) was the incorporation of assessment into instruction and away from formulaic and "add-on" assessments to more integrated, holistic, and imbedded assessments of reading ability. Assessment is best used when it helps to instruct. The researchers felt that summaries were a way of assessing reading ability and were not meant to be a product to be assessed in themselves. They realized that for some teachers assessment and instruction were entirely separate; teachers were using the performance-

based assessments as an extra, added-on activity, not imbedded in their day-to-day instructional practice. In Hiebert and Davinroys' paper, "An Examination of Teachers' Thinking about Assessment of Expository Text" (1994), the authors illustrate how introduction to an unfamiliar genre of literacy --expository text-- resulted in changes in thinking by teachers as they gained new understanding of summarizing as being a "more generic process" rather than formulaic. For some teachers, changes were made in how summaries were used, with teachers using a wider variety of assessments with differing formats and purposes.

The CRESST technical reports provide comprehensive information about lessons learned by researchers. Other papers and reports summarize the results of the research endeavor. Five propositions found in the Borko/Flexer paper, entitled "Teachers' Developing Ideas and Practices about Mathematics Performance Assessment: Successes, Stumbling Blocks, and Implications for Professional Development" (1995), provide experienced advice to others conducting similar research. The first proposition advises researchers to situate learning about any new techniques in assessment and instruction in the classroom. Teachers learn best about the benefits of new techniques by discussing the technique, using it, witnessing the results, and then discussing what they have learned. They also learn about assessment and instruction when they discuss what they have learned in groups (the second proposition); new ideas are often worked out and constructed by teachers working together.

The third proposition looks at the role of the researcher in "scaffolding" teachers' understanding and skills. Researchers must walk a fine line between telling teachers what to do and giving up on providing any guidance. Questions, explanations, and suggestions are examples of

ways in which discussion can be guided to the teachers' level of understanding and provide a means for teachers to discover and explore new skills and ideas. The fourth and fifth propositions are discussed more in depth in the present paper.

The fourth proposition relates to teacher beliefs. Some teachers came to the study with beliefs "incompatible with the intentions of the staff development team." Borko thought that if these beliefs go unchallenged teachers are "likely to ignore new ideas or inappropriately assimilate them into their existing practices." Because of agreements made by the CRESST team to avoid direct challenges to beliefs, the researchers thought that opportunities were lost "to help [teachers] think in new ways about mathematics assessment and instruction."

The fifth proposition relates to time. Teachers found that the scoring the new assessments and keeping observational records was taking time away from instruction and were a burden to implement. The researchers recommended that staff development efforts take place over a long period of time and provide released time for teachers so that they can become proficient users of performance assessment. Additionally, researchers recommended prioritizing research goals ane that researchers not be afraid to rethink goals in the face of time pressures.

The present paper discusses several propositions in more detail and provides additional lessons learned by asking the participating teachers their views. Teachers also reflected on the impact on their teaching and assessment, and what benefit they perceived from participating. An interview with Hilda Borko illuminated some issues and framed the concerns of the teachers.

Difficulties Encountered

CRESST research about performance-based assessment was conducted in three participating schools in the same school district during the 1992-1993 school year. Three other schools in the same district were used as controls for comparison of outcome student outcome measures (Shepard et. al, 1994). During the 1993-1994 school year, two of the three schools continued to participate, and during the 1994-1995 school year case-studies were conducted of two teachers in one school.

Three teachers, one from each participating school, were interviewed. Rhonda, Sally, and Beth (not their real names) participated in the study for varying amounts of time. Before the project began, CRESST investigators contacted district administrators. School principals offered teachers the opportunity to participate. Some teachers felt that they had not truly volunteered for the project because of logistical considerations, and in some cases, pressure from administrators who wanted teachers to participate.

> Although all fourteen participating teachers were technically volunteers, some were less enthusiastic than others to engage in the project. Some of the teachers who were part of the original application process changed grade levels or schools and were replaced by other teachers who found themselves involved in a project for which they had not volunteered: others may have been "strongly encouraged" to volunteer by the principal or other teachers in the school (Borko, 1995).

Rhonda was one of the teachers who felt that the principal of her school had "highly encouraged" her to join the project. She told of her experience:

I believe that CU approached the administration building first, and because of the reputation that [our school] had, the assistant superintendent at the Ad building approached the principal at the school and felt very strongly that this would be something good for [our school] to do... It was highly encouraged that we would do it. So we did it. (intR: A4)

At other schools, teachers felt that they had freely volunteered. Beth and Sally did not feel any pressure to participate and felt that they had chosen to join the project freely. Beth said that one CRESST partner told the teachers about the study. At both schools the teachers involved in the study discussed the project among themselves and made the decision to participate.

One reason for resistance may have been the feeling that the demands of the project were not presented in a clear manner during the recruiting process. Rhonda felt that the teachers at her school were not given a fully informed picture of the amount of work and time needed to participate. She said that her school was participating in several different projects at the same time, and that the administration and the project members didn't seem to understand how much work was involved for those participating.

I felt that the amount of work that would be involved wasn't something as clear from the Ad building, and the people running the program. I don't think they were trying to hide anything from us, I think because we're in the classroom we realize how much work is involved, and therefore that piece on their part was a little bit

unclear... I don't feel that it was done on purpose, but I don't think they realized how involved this project would be as far as taking time, and as teachers we really did know that, so we were really on the fence [about joining the project] (intR:A4).

However, both Beth and Sally felt that they had a clear idea of the amount of time and effort they would be expected to spend on project activities. Neither felt that there had been problems communicating with project members. Recruiting unwilling and resistant participants led to some difficulties. Presented with a heterogeneous group of teachers, some of whom were unwilling or resistant to the agenda, researchers reexamined their goals and methods. Hilda Borko explained:

There was more variability in the teachers than we expected, because we had asked for volunteers, and some of the people it turned out their principal volunteered them, or the teachers from the previous year volunteered them, or the teaching staff changed after the agreements had been made. So I think we found some people who were less interested in, and to some extent more resistant than we anticipated, and we found some teachers who didn't make some of the changes we hoped to see. (intBk:77)

Because of the agreed-upon restrictions put in place before the study began, researchers could not directly challenge the beliefs and instructional practices of the teachers. When the researchers found out that the teachers were not following the district curriculum (e.g., using assessment that promoted problem solving), their task was made more difficult. The difficulties arising from this situation are discussed in detail by Borko et. al (1995). Borko noted the incongruence between

teacher and researcher beliefs and suggested that researchers faced with this situation should make agreements with teachers to examine their beliefs about assessment and instructional practices before the research begins.

> *Once we got into the study, we discovered that despite some of the ways that we selected teachers we had made assumptions on the basis of our selection process, and some of those assumptions were not borne out. For example, the match between their programs, and the district curriculum framework was not borne out. So, knowing that in retrospect we would've done something differently, we would've started out up front with some agreements to look at their instruction too. We also found that some of their beliefs didn't match what we expected, yet...because we had assumed that their instructional goals were similar to the district curriculum framework. That turned out for some people not to be the case, but we had no agreements to explicitly address their beliefs, to look at their beliefs, to have them look at their beliefs. So, after the fact, if I were to do it again I would have beliefs be a part of intervention agreements up front in addition to... assessment practices. (HB:49)*

Rhonda's school was an example of the situation in which teachers held incongruent beliefs and practices. At first, teachers at her school were resistant to suggestions from the research team to examine their assessment practices. She remembered the teachers telling the project members "point blank" the third week of the study, "We like our math book. Why do we need to stay here until five or six o'clock and reinvent the wheel?" Because of the incongruence between the beliefs of the project

members and the teacher's beliefs, project members turned their attention to the teacher's beliefs, but didn't directly challenge them.

Rhonda explained how researchers let the teachers come to their own conclusions about their assessment practices. One method she remembered the researchers using was having the teachers compare their own lesson plans with the official district curriculum plan.

> *They were very patient with us,. They made us think... They led us down the pathway in a very tactful way and asked us: Is your math book meeting your districts curriculum? Are you teaching your curriculum, or you just teaching HBJ? .... We matched the district curricula to the math book, and we found out there were all these holes.(intR:K3)*

At the other schools, the assumptions made by the researchers were more in line with expectations. Sally felt that the project "supported the way she taught and assessed" and that she shared many beliefs with project members before the project began. Beth also felt that her beliefs, and some of her assessment practices were congruent with those of the researchers before the project began.

Insuring that teachers hold beliefs that are congruent with those of the researchers is difficult, and may be unnecessary. Borko believed that there was no realistic way (i.e. survey or preliminary interview) of determining the beliefs of teachers before the project began. She added that since the teachers at the participating schools were supposed to be following curriculum guidelines congruent with the research agenda, the assumption that teachers had congruent beliefs was not out of line. She advised that instead of assuming beliefs and practices are congruent with

expectations, it would be better to build in attention to beliefs and instructional practices in preliminary agreements.

Successfully participating in the research project demanded a great amount of time and effort on the part of the teachers. In the CRESST technical report, the authors state: "All the teachers found the additional work in the project burdensome in the fall, and by Thanksgiving, they were feeling overwhelmed" (Flexer, et. al, 1994, p. 16). In response to this complaint, the workload was decreased.

Rhonda said that the teachers were often in the building until six or six-thirty at night working on the project and that her personal and family life suffered during the first year. Sally and Beth reported similar time pressures. While each teacher said that the heavy time commitment was necessary to successful implementation, they were happy that the commitment had paid off in terms of better assessment and teaching. Borko agreed that a significant time commitment is necessary for teachers who participate, but added that project members found ways to mitigate the harsher effects of the time commitment. None of the teachers interviewed thought that the project had caused them to teach less effectively because of time pressure, and all felt that they knew beforehand the amount of work that the project entailed.

"Comp time" was perceived as an important way to mitigate the effects of the heavy time commitment. Each teacher said that the small amount of time set aside during the school day (usually four hours a month) helped them meet their goals, and each teacher said that successful participation in the project would have been impossible without it. Rhonda remembered that teachers on her team had no trouble filling their comp time, and she said that teachers would work "straight

through" their allotted time without breaks. Her team could have easily used a full day to do CRESST work.

Sally and Beth gave similar opinions and said that more comp time was needed to complete this type of project. Sally and Rhonda added that several research or professional development projects were running concurrently at their schools and that this fact should be taken into consideration when planning for any project. Borko agreed that comp time or inservices were valuable for any research or professional development, and encouraged those conducting similar research to take this into account.

The project had been planned for one year, but was extended at two of the three schools. Borko felt that this extension was necessary for teachers to become comfortable with the changes brought about, and advised other researchers conducting the same type of research to consider making this type of longitudinal time commitment.

> *I would definitely build in at least that second year because it became clear that as people began to make changes in the first year, but they are not fully able to incorporate them into a sort of yearly curriculum until the second year, at best! For some of them it was the third year before they really started to get comfortable. So had teachers not been willing to stay with us an extra year, there's a lot we wouldn't have learned. And, you know, one out of the three schools didn't. That would be the main change.(HB:122)*

The CRESST team learned about the assessment and instructional practices of the participating teachers in a variety of ways. During the first year, workshops were held where teachers discussed their

assessment practices with the researchers. Teachers were also interviewed on a one-on-one basis. During the second year, researchers added observations of classes to their protocol and talked to teachers after the classes. Borko said that observation was a valuable tool for researchers, and felt strongly depending upon discussion as the sole means of collecting information was inadequate for understanding the changes that were occurring. Using observation in the classroom as a form of triangulation helped avoid "cross communication."

> *I think it's really important to go into classrooms and to talk to teachers in a face to face situation, because it's very easy to think you know what teachers mean by something, but until you see how they enact it in practice you don't know what they mean by it. Conversely, it's easy for teachers to think they understand what we're asking, but if they have a different definition of what counts as 'problem solving," or a different definition of what counts as kids exploring through manipulatives, then everyone could be acting in good faith and they would still be talking across each other, instead of to each other. (HB:152)*

Using observation in the classroom is also an important way to develop better communication and a mentoring relationship with the teachers.

> *We found by spending some time in their classrooms,that being in their classrooms, observing in their classrooms, gave us valuable data about what in fact their practices looked like, rather than only have what they reported and what they brought in. We also found*

that by being in their classrooms we could talk easily to them afterwards, and give them suggestions that were concrete and were naturally flowing from what we saw, and I think that's a really powerful intervention. It's a labor and resource intensive intervention, but I think that I would build that in next time. (HB:65)

Finally, observation was a way of controlling socially desirable responses on the part of the teachers. While Borko felt that this wasn't a serious problem for the CRESST project, classroom observation is a way of controlling for teachers who offer responses that they think the researchers want to hear.

How Alternative Assessment Benefited Teachers

The results of the CRESST project and its impact on the assessment practices of teachers can be found in the CRESST publications. The question of how teachers benefited is partially addressed in these publications. We sought to learn more about the impact on teachers and if the teachers themselves thought the results to be positive and worth the effort. Borko offered her interpretation how they benefited.

There were two groups of people I would really say that benefited. One group who already had a belief system and some ideas that were real compatible with ours, and they could sort of take this and run. Every activity, every assessment tool, every idea we had, it was easy to take what we had and go with it. I would say they changed, but not the most. I think the people who changed the most are people who really did some shifts in their assessment and teaching, that were pretty dramatic toward performance assessment, toward more student centered assessment and instruction, toward more

integration of problem solving, and well-defined problems in math (HB: 103).

All three teachers interviewed used some type of performance-based assessment before the project began, but both Beth and Sally expressed beliefs and reported practices that were more congruent with those of the researchers than Rhonda. Beth said that her school had already emphasized the use of lit conferences and portfolio assessment before and had implemented district guidelines emphasizing use of rubrics. However, the teachers all said that they depended primarily on informal observation to know how the students were performing, and seemed to take assessment for granted. Rhonda was more traditional in her assessment practices. She remembered "keeping a lot of information in my head," and depending on tests in the math book to assess her students' math skills. Of the three teachers, Rhonda reported making the greatest changes in her assessment practices.

The teachers reported a number of beneficial effects from the time they spent discussing assessment, practicing what they discussed in their classroom, and reflecting on their experiences in workshops and interviews. All teachers felt strongly that the project was worthwhile, and that they would participate again. Rhonda, who was at first resistant, thought the project had "made me a better teacher." Sally said she had "experienced a lot of personal growth.". Beth said that the time she had spent had "got my mind going." These and other positive comments indicate that the project was perceived as successful and worthwhile for the teachers.

Teachers gave a variety of reasons why performance-based assessments helped them. These reasons are the same as those in the

literature about performance assessment: an emphasis on problem solving on the part of students, student involvement in their own assessments, and greater ability on the part of teachers to troubleshoot and diagnose the problems their students experience. One benefit was a new awareness about assessment. Sally said that participation had provided her with a conscious awareness of assessment issues and of the question, "How do I know what my students know?"

> *I remember at the beginning of the year when [the teachers and the project members] first started talking. The question was how will you know, how will kids demonstrate that they know this to you? I mean I can't believe this, but at first I said, "I don't know." So I think just being tuned in to it, always being a conscious effort, a conscious decision, how will I know what my kids know? What do they need to do to prove to me their understanding? (intS: E26)*

One of the primary perceived benefits was diagnostic utility for assessing student achievement and thought processes. Teachers said that they knew more about their students from using assessment that incorporated and displayed student explanations in math, and the use of running records which kept track of the number and type of errors students made in reading. Sally expressed the satisfaction one student felt when he noticed the progress he had made in reading, evident from the use of a running record.

> *...you always had a gut feeling how kids were reading and where their errors were. But with the running records you have documentation... We had a student who was significantly below grade level in reading.*

At the beginning of 3rd grade I had done a running record on him and a primer book -- it was the I Can Read books, and that was in October-- and he read a page and a half and just struggled and I stopped because of the pain on his face... and I wrote down all that he was reading and what I observed and then I gave him the same piece in January and he blew right through it. So right then I went and got his October record and put it in front of him and we looked at all the marks that I had made and how far he had gotten and I said to him, I said, what does this tell you? And he just beamed, and said, I'm getting better. And that was, I would never forget that day and had I not had that document he would have had to take my word for it. Look at how much better you're doing and it's hard for a kid to remember. That way when you can see it on paper, I think that was really helpful. (intS:C64)

The ability to diagnose and remediate difficulties in reading and math from the use of performance-based assessment was mentioned by all three teachers. Rhonda displayed a sample of performance assessment related to the greater-than and less-than concept in math. She explained the advantages:

And when there was a mistake I could understand what the mistake was. I could tell if they could- like this one, "one less" I mean, she just wasn't thinking, and I could tell that from this paper. I know when I talked to her mom about it, I said this answer here just shows that she wasn't as close attention as she should've been. (INTR: C20)

Beth thought that she no longer had to take home "a huge stack of papers" every night to learn about the progress of her students. A closely related benefit was the ability to involve students in their own assessment. As students participate in producing a rubric for a writing summary or math problem, they must think about what criteria make for good writing or a good explanation in math. The teachers thought that the participation of students in the assessment process was one of the positive impacts. Sally related how, as students became involved in creating rubrics, they gained a better idea of why they received the grades they did.

> *I think the other real valuable thing that came out of the project was a lot of student self assessment. You know, when we were writing rubrics together, one of the things was that we talked about it. We felt like kids that were performing low often weren't aware that they were low or didn't know why they were low. [They said] "I know I don't get this, I know I'm not doing a good job, but I don't know what a good job is. I've never been able to do it and all I see is that is I'm getting these low grades." It was like, I can't believe we were doing this to kids, not ever telling them, you know, what is missing, so that you can have a "4"." So I think [the benefit] for kids is that struggle, the self-evaluation and the sharing of rubrics. I think sometimes those kids thought their grades just came out of a hat! (intS: E28)*

Beth also thought that self-evaluation and the participation of students in their own assessment was an important benefit.

... it gives them some real goals for their learning. If they know what they're doing they can see what they need to do. Our district has a writing rubric that we use, and I usually go over that on the overhead with the kids a few times a year." This is what's expected, can you think of anything else you do well that we should put on the rubric?" So I kind of have the district rubric, and then I'll talk to them about what else they think should be on there. And then, you know, talk to them about the numbers and have them score each others' pieces to kind of see where they think they would fit on the rubric. So it's a real interactive situation. (intB:E23)

Changes in assessment also led to changes in instruction. Rhonda thought that before the project began she did a lot of "feeding information" to students. Through emphasis on problem solving, she now feels more comfortable letting students figure out lessons on their own and acting more as "a facilitator" than a traditional teacher "spoon-feeding" information to the students.

I think all day long, because of the project, and because of the way that I've changed, is that nothing goes on in the room, hopefully, that's just rote, they just do it. They understand why they're doing everything, and they realize that there isn't going to be anything given out that they're not going to have to think about. They can't just circle answers. They're going to have to think through an activity (intR: E10).

Sally also believed that her class is more focused on "problem solving, reasoning, explaining, and thinking" than before, and that the incorporation of running records provided new ways to teach language

arts. Because of the diagnostic utility of performance-based assessments, teachers thought that they were better able to communicate with parents. Having detailed and specific documentation of student progress lets teachers give parents a better idea of not only how well students are faring, but exactly why and where they are having difficulty. Rhonda said that since the project ended she has developed summaries of student work that she uses to tell parents about their children's achievement.

What Teachers Said About Changes Assessment Practices

How teachers assess their students changed. While all teachers report d that informal observation was their primary method of "knowing what their students know," they also reported using a variety of performance-based assessments that they developed. They varied in use of different types. Rhonda reported using many different assessments in math, but only a few reading or language arts assessments. Sally and Beth both said that they used running records and written summaries extensively, while Rhonda almost never used running records, and only used summaries occasionally because "the students hate them."

In math, Rhonda's students do a variety of both paper and pencil and performance tasks for assessing the math comprehension, calculation, and understanding. Rhonda gives her third-grade students a "problem of the day" to assess their problem solving skills. One such problem presented students with a picture of blocks; students figured out how many cubes were shown in a picture. While many students only counted the cubes that were showing, the difficulty in the problem was perceiving that some of the cubes were hiding underneath the visible cubes. What makes this a form of "alternative assessment" is that students must back-up their

answers with an explanation. and this explanation is part of the grade the student receives.

Rhonda debriefs each problem with an explanatory poster and a discussion, and tries to make sure that all of the students understand the concepts underlying each problem. Rhonda also has a variety of exercises that students perform during class that help her understand her students' achievement. A typical exercise presents students with three numbers -- a "family of facts"-- students must explain how the numbers fit together (i.e., the numbers 5,11,16). Other assessments are more like traditional word problems, with the difference that part of the score on the assessment depends on how well the student explains their answers. Scores are judged by the criteria from a rubric.

Performance-based assessment also involves performing tasks. In Sally's class, students built and designed their own playground area, a long-term task that involved planning, measurement, and drawing. She said that she has shifted the emphasis in her math instruction to more long-term, multi-stage problems, but that the means of assessing these tasks is still anecdotal. In Rhonda's class, students do a "polyhedran" problem where students in groups measure an edge of a cube and then figure out the total measurements for all edges.

In reading, both Sally and Beth use running records. Running records are a document of the errors students make in reading. Examination of the record demonstrates progress and points out the difficulties a student is experiencing. Both teachers felt that these assessments were valuable and had changed the way they taught. Beth and Sally said that they have students write summaries about what they had read, and they grade these summaries using a rubric that the students developed with the teacher.

While Rhonda does not use running records, and only uses formal summaries with a rubric occasionally, she did use reading assessments that asked students to summarize the plot and characters of different chapters they were reading. She also used a "double diary" in which students write events from a story they're reading and then gave events from their own lives that paralleled those of the story characters'.

The teachers use a variety of other techniques. Each uses some form of portfolio assessment, and each keeps anecdotal notes about their students, though Rhonda admitted that she didn't keep notes as often as she did during the project. Sally said that she used a formal observational checklist in math, but both Rhonda and Beth said that a checklist was impractical and awkward, a result attested to in the Borko/Flexer paper. Each teacher met with students on a one-to-one basis to assess progress, though only Rhonda said that this was a formal part of her daily assessment practice.

While each teacher used a variety of assessment practices, most still used traditional assessments as well. Rhonda gives her students five straight computational problems in her math class every day, a weekly spelling test, and some chapter tests in reading. However, she avoids the math textbook and has used it only four times during the first three months of the school year. Sally uses paper and pencil computational worksheets in math, a weekly spelling test ('for the parents- I don't even grade it"), and some traditional reading skill exercises. Beth also uses some traditional paper and pencil worksheets. All three teachers have given up on heavy dependence on textbook pre- and post tests and say that their assessment emphasizes problem solving and comprehension more than basic skills.

None of the three teachers looks forward to giving the standardized multiple choice tests that the district administers to their students. Rhonda said that preparing her students for the multiple choice test "makes her angry." Sally had a similar negative opinion about multiple choice tests. She remembered the changes in her classroom brought about by the yearly testing:

> *... that's why we really wanted to be part of the project because we were so frustrated with teaching one way and then every Spring that CTBS test came around, and that was ridiculous! Our kids have never been tested in reading this way. We have never given them two paragraphs and have them answer 10 questions multiple choice to assess their reading. Or in math -- we never timed them on a bunch of computations or had them do multiple choice in math.(intS:A14)*

While teachers must still prepare for the testing, they say they have a more sophisticated understanding of assessment and testing than before. Rhonda said that her district had implemented a multiple-choice test that was meant to measure how well teachers were implementing the curriculum. She said the math portion had some positive characteristics, such as a minimum of straight computation, and an emphasis on conceptual understanding. She praised the test for allowing students to use calculators. She mentioned other attributes of good assessment, such as avoiding straight percentage scores for giving grades, and not administering timed tests.

Other considerations

The three teachers said they had benefited from the resources provided by the project. The resource they found the most useful were the

project members themselves. Beth thought that being able to consult members of the CRESST team who had extensive knowledge of the latest research helped her find and develop the assessments she needed. Rhonda felt that "she had her own private math tutor" and was comfortable asking for help with assessment and instruction from all of the team members.

Much help given to teachers was advice on where and how to find the resources they needed. CRESST provided articles about assessment and the "skeletons" of assessments themselves. Sally said that the project provided resources such as Marilyn Burns activities, materials from the Lane County Math project, and other written materials. CRESST also provided notebooks of readings that helped teachers plan language arts activities. Beth remembered receiving tapes for instruction, conducting running records, and project members giving her instruction in construction of rubrics.

It is difficult to judge the extent to which the teachers who participated in the CRESST study influenced the other teachers at their schools, but it does not seem to have been a great deal. One teacher has changed schools, while another was participating in a new project that had similar features to the CRESST project. Rhonda said that teachers at her school had changed little since the project began. Recently, her school was asked to develop performance-based assessments to meet state requirements. Rhonda described the state of understanding this way:

> *We gave a presentation twice to the faculty, and they were really interested in it, and they really understood it, and last year with the nine or ten mathematical strands in the standards, we were asked by the administration to come up with a performance based assessment*

for each of those strands. This literally threw our faculty into a tizzy. They had no idea what to do, but in third grade we whipped ours out pretty quickly, and we're pretty pleased with them. The rest of our staff -- it's a philosophy, and you can't just get it over night, and we worked with those professors for two years. And when we make our presentations, even though they are really interested, and they ask about the problem solving,they still teach problem solving like it's a three week unit...

..And if we're in the workroom this year, we still see all the chapter pre- and post tests. So, even though they're really interested, and complimentary of it, there hasn't been a big spill over. (intR: RAW167).

Conclusion

Greater attention to standards may force many school districts to become more involved in promoting a curriculum that encourages problem-solving and higher-order thinking. However, this change likely will be gradual and gain momentum only as it works its way from teacher education programs to practicing teachers. Professional development programs such as this one can provide valuable insights into how this can be accomplished, but obviously cannot bring about wide-spread changes through these efforts alone.

The three teachers interviewed felt strongly that the project was worthwhile, and all three said they would "go back and do it again" if given the chance. While some may be implementing performance assessment in an algorithmic, or formulaic manner, all three teachers have adapted their classroom practices in their own way to take advantage of a new type of assessment. Along with this has come a new way of thinking about what

students can do: the teachers emphasized problem solving and wanted their students to understand why they get the grades they do.

Interventions involving extensive program development are complex because researchers must deal with entrenched instructional practices and school cultures that may not be congruent with the R&D agenda. Taking differences in beliefs between teachers and researchers into account should be part of professional development. Time is also important: researchers should plan their intervention so teachers have enough time to work on the implementation during the school year, and should not expect dramatic changes over the course of only one year. Finally, it should be emphasized that professional development works best when researchers can discuss changes needed, but also observe and mentor teachers in the classroom as they implement new techniques.

References

Davinroy, K. H., Hiebert, E. H. (1994). An Examination of Teachers' Thinking About Expository Text. (Unpublished).

Hiebert, E. F., Davinroy, K. H. (1993). Dilemmas and Issues in Implementing Classroom-Based Assessments for Literacy. Presented to AERA, April 13, 1994. CSE Technical Report 365.

Borko, H., Mayfield, V., Marion, S., Flexer, R., Cumbo, K. (1995). Teachers' Developing Ideas and Practices about Mathematics Performance Assessment: Successes, Stumbling Blocks, and Implications for Professional Development. Presented to AERA, April1995.

Flexer, R., Cumbo, K., Borko, H., Mayfield, V., Marion, S. (1994). How "Messing About" with Performance Assessment in Mathematics Affects What Happens in Classrooms. Presented to AERA April 1995. CSE Technical Report 396.

Shepard, S., Flexer, R., Hiebert, E., Marion, S., Mayfield, V., Weston, T. (1994). Effects of Introducing Classroom Performance Assessment on Student Learning. Presented at AERA, April, 1994.CSE Technical Report 394

Chapter 6

Conclusions about CRESST Impact

Ernest R. House

Evaluating the impact and influence of R&D work entails special problems. One is that the impact of research and development may be delayed for many years or even decades. Another is that the pathway of influence from the R&D work to practice is uncertain, so that one cannot predict precisely how the R&D findings and products might reach their ultimate destination, or even what form they might take if they do. For example, R&D findings and products in the hard sciences often are combined with other work to create new ideas or products altogether, and products not anticipated by the original developers.

One approach to such uncertainty is to divide the effects of the R&D into anticipated and unanticipated effects. Although focusing the evaluation on the anticipated may underestimate the influence of the R&D work by not accounting for surprises and long-term, serendipitous effects, it permits shorter-term, intermediate evaluation of impact, albeit conclusions which must be qualified because they are incomplete. Unanticipated effects might be studied later, probably by retrospective case studies, to complete the picture.

There is also the question of the quality of the R&D work. For the most part we have attempted to judge the impact, rather than the quality of the R&D. Quality is certainly of foremost importance but it has been handled traditionally via peer review procedures, however restricted in rigor these may be. CRESST work has been subjected to any number of peer reviews since it has been funded by the US Department of Education. Furthermore, the articles published in journals have been subjected to

peer review for the most part. Nonetheless, comments about the perceived quality of the work appear in these studies often since it is a major reason given for influence.

In this evaluation of the impact, we have assessed the influence of the CRESST R&D work on several populations over a five-year period. Although the ultimate aim of CRESST is to improve the testing of students in such a way as to enhance learning, mainly through techniques known as "alternative assessment," we would not expect CRESST efforts to have resulted in widespread application in classrooms or in increased student achievement throughout the country after only five years. The ultimate impact sought is too grand to be accomplished in such a relatively short time period.

The primary question for this evaluation then is what measurable impact, influence, or effect the R&D center has had during its existence these past five years. We use the terms impact, influence, and effect almost synonymously, though we recognize there are more distant goals to be achieved ultimately. We assessed CRESST influence on researchers, anticipating that CRESST must influence the educational measurement community to reach its goals; on test directors of school districts, anticipating that these are gatekeepers to the student assessments that are conducted in districts and states; on users of CRESST products, anticipating that these products must work well if CRESST work is to be successfully implemented, and on a few teachers, anticipating the issues and problems that might be encountered when alternative assessment is implemented in many classrooms.

To judge influence on the educational measurement research community, we conducted analyses of CRESST publications and citations,

including how many publications were produced, the status of the journals in which they were published, how often these publications were cited and in which journals, and how citations were used in the context of the article of citation. According to all these indicators, CRESST researchers had a very substantial impact on the educational measurement community. They produced a large number of publications in the highest status research journals and had their work cited frequently by other researchers in ways central to the development of the ideas in the articles of citation. CRESST also published articles in a number of practitioner journals as well.

For example, the core group of CRESST researchers produced 90 articles, books, book chapters, and technical reports that were cited 424 times between 1990 and 1995, or 4.7 times on average. This is a substantial number by almost any standard. Furthermore, the articles were published in many of the highest status journals in the field, indicating acceptance by peer review and access to the leading scholars. When self-citations (17%) and CRESST partner citations (24%) were removed, the majority of citations (59%) were by researchers not connected with CRESST. Of course, these citations were not normally distributed. A few publications garnered most of the citations.

On a three-point rating system of journal status, the CRESST publications rated 2.3 on average. This high status ranking was achieved in spite of the fact that many CRESST researchers published articles in lower status practitioner journals in order to influence practice, thus bringing down their overall scores.

Use of the articles was also analyzed. For example, the Linn, Baker, Dunbar (1991) article has the most citations among key CRESST

publications, with an average citation rate of 13.1 per year. On a three point use scale from least to most important use, it's average was 1.8. In general, this number reflected how most CRESST research was used in citations. A use rating of "3" indicates that the work was critical to the article in which it was cited, and "1" indicates that another publication could have been used in its place, even though it was cited. So centrality of use was substantial for the nominated CRESST articles analyzed.

To determine the boundary of CRESST influence on the research community, an ERIC search for 1994 and 1995 discovered 54 performance assessment articles, 10 of which were published by CRESST partners, or 18.5% of the total, a substantial portion to emanate from one research program. In an examination of 35 of these articles CRESST research was cited in 90 times in CRESST authored publications and 42 times in 26 non-CRESST articles. Most CRESST articles were in the highest status journals. All in all, the evidence is extensive that CRESST has had a major impact on the measurement research community.

One limitation to this bibliometric study is that there are no comparable groups of researchers to compare to those in CRESST, other than those in natural science or other fields in which the publication and citation practices are different. However, even without such comparisons, the volume of publications, citations, and uses is so high that influence and impact on this particular research community are clear and unmistakable. One would presume that alternative assessment could not advance into practice without the endorsement of the major relevant research community.

The second community of influence was that of test directors, publishers, and others who serve as the gatekeepers to district and state

assessment procedures. A national survey of this population revealed the strong influence of CRESST R&D on them as well. The test directors were convinced that alternative assessment was important, a significant improvement in assessment procedures, and that CRESST was a major and highly valued source of information on the topic. Not only did the directors agree that CRESST was a major influence on their thinking and their decisions to use alternative assessments but they also lauded CRESST for the high quality and objectivity of its work. CRESST was perceived as an intellectual resource on which the directors could rely.

As a group, the test directors were open to alternative assessment techniques, with writing assessments, performance tasks, and portfolios being the most popular, and exhibitions, experiments, and oral exams the least popular. These alternative forms of testing were used at a high rate, albeit at a rate less than that of traditional, standardized tests (not surprisingly). Of course, a few directors did not like alternative assessment at all, mostly because of its perceived lack of validity and reliability.

There are important qualifications to this finding. First, the test directors accepted alternative assessment techniques only as a supplement, not a replacement, for traditional standardized achievement testing. For the most part they saw alternative assessment as useful at the classroom level to improve teaching. It was not seen as useful for accountability at the district or state level. Traditional achievement testing was perceived as best for that. This surprisingly positive attitude of test directors towards alternative assessment probably would change if they were forced to choose between alternative and traditional assessments.

The test directors saw CRESST as an important and influential source of information on alternative assessment. About 30% of respondents saw CRESST as very useful, 31% as useful, and 24% as somewhat useful. Only 4% saw CRESST as not useful at all as a source of information. In open-ended comments they were extremely laudatory about the high quality and objectivity of CRESST information and research on the topic, perceiving CRESST as a valued and reliable source in an area in which much information is suspect. CRESST personnel received high marks for competence.

Most directors' previous experience with alternative assessment has been with writing assessments, though alternative assessment is a recent experience (the last two years) for half. The directors also relied on many other sources of information other than CRESST, so their attitude cannot be attributed solely to any one source. Impersonal sources of information, such as journal articles, seemed to dominate their contacts with CRESST, though about one-quarter had personal contact of some kind. One can conclude from this survey that CRESST has had a highly significant and continuing impact on this important gatekeeper community of test directors, whose acceptance one would imagine as critical to any widespread application of alternative assessment techniques.

A third study examined the influence of two CRESST products to see both how and how widely the products were used. Products are closer to the practitioner community than journal articles. CRESST was asked to nominate two of their best products, and we attempted to track how these products were used in the field. Admittedly, these tracer studies have a positive bias because we asked CRESST to nominate two of their best products and suggest the names of those professionals who had made use

of these products. CRESST would not likely nominate poor products or suggest the names of those who didn't make use of them. One might think of these tracer studies as "best case" scenarios.

The first product was the book, A Practical Guide by Herman, Aschbacher, and Winter. This publication was distributed to 90,000 members of the Association of Supervision and Curriculum Development (ASCD) through their regular publication list. ASCD is an organization of curriculum directors and others in school districts who are responsible administratively for curriculum matters at the district level. Clearly, this is a huge distribution for any product, using the curriculum network of ASCD. CRESST also distributed another 44,000 through other sources.

Telephone interviews with those who have used the book extensively reveal a high regard for the product's quality and usefulness. Mostly, the book was used for training teachers and administrators in alternative assessment techniques, and such a product was badly needed in the field, many respondents reported. The state of Illinois has used the book extensively as part of school planning processes. Most users thought the book covered the essential topics in easy to understand language and was put together so individual chapters could be used rather than the entire book.

Using the existing ASCD network helped the distribution considerably. Most said they had been looking for something on alternative assessment to use before discovering the book. Often, use started with a state department of education initiative, which stimulated districts. Word of mouth was a favorite dissemination pathway, with most users saying they had recommended the book to 25 to 100 people. Brevity and simplicity were perceived positive attributes.

The other product, a content assessment model based on cognitive psychology, has been used by Hawaii and Los Angeles schools. However, these and other projects in Missouri and Washington are not far enough along to evaluate definitively. Users have tried to develop assessments based on the model, but, in general, the model requires complex implementation and adaptation. Its success cannot be judged at this time. In the case of both the book and the model, users were already aware of alternative assessment but needed tools for implementing it.

The final study was an attempt to assess impact on teachers, the ultimate group who must implement alternative assessments eventually. At this point in time relatively few teachers across the country have made use of these techniques. The techniques are still under development and too new to be widely distributed. Hence, there was no sense in conducting a national survey that would meaningfully represent teachers across the country. Most would have had no experience with alternative assessment. Yet teachers are the most critical group of all (except for students), and the ultimate success of alternative assessment must rest with their eventual acceptance and use.

To address these problems we examined one of CRESST's pilot projects. CRESST has undertaken a few pilot studies in which a small number of teachers were helped to develop alternative assessment measures for their classrooms. We examined the documents produced by these pilot projects to estimate the eventual benefits and problems that teachers across the country were likely to encounter as they try to implement these new measures. We also interviewed a few participating teachers independently after the pilot project was over to obtain their retrospective views.

In general, participating teachers adopted the techniques that fit their underlying beliefs, but not those that did not. Pre-existing beliefs of individual teachers towards instruction and assessment and their reliance on text books turned out to be significant factors in the implementation. The development also took a great deal of time, and some teachers thought that the effort detracted from time spent on instruction, necessitating a reduction in workload eventually. "Comp time" amounting to a full day a month would have helped ease the extra burden. Also, other projects on-going at the schools, unrelated to CRESST, made things more difficult.

A second year of project participation was needed, and even then only by the third year did teachers feel comfortable with the new ideas. Nonetheless, the participating teachers thought they had gained by improving the performance assessment of students, involving students in their own assessment, and diagnosing the problems students faced, which included being able to discuss student performance better with parents. In general, assessement was more integrated with instruction. Participating teachers varied individually in what they did and how they did it, including their acceptance of the new ideas. Nor did the new ideas spread to the other faculty in the schools over the three year period.

Finally, according to information from the Lexis/Nexis news service, since 1990 there has been a rapidly increasing number of articles on the topic of alternative assessment. In 1990 there were five articles, two in the New York Times, two in other major newspapers, and one in an educational publication. In 1994 there were 85 articles, 42 in regional or local papers, 22 in educational publications, 18 in mass circulation magazines, and 11 covering Congressional testimony. First, major

newspapers recorded the trend, picked up later by local and regional papers, and followed by Congressional and legislative sources. Most articles described new assessment policies and controversy over their introduction. Articles in the major papers explicated the trend towards alternative assessment. One of the earliest appeared in the Los Angeles Times in March, 1981, an interview with Eva Baker, co-director of CRESST. Although CRESST was mentioned by name in only a few articles, this underestimates its influence since it works through intermediaries.

There are at least two significant limitations to this evaluation as originally conceived. First, we did not ascertain CRESST influence on policy makers. There are two reasons. When we queried CRESST researchers as to the policy makers with whom they had worked, their response was not systematic. In answer to a survey, they responded by describing work in which impact was assumed, by stating future impact they hoped for, by describing how they had disseminated information, by citing the interest of other reseachers, or by listing specific instances of actions caused by their work, each responding in a different way.

Defining a set of policy makers whom we could interview would have required considerably more effort, starting with interviews with CRESST researchers themselves, including

1) a clear statement of audiences the researcher wanted to impact,

2) what evidence they themselves believed constituted impact,

3) descriptions of networks through which impact might occur,

4) follow up interviews with key people identified.

Furthermore, surveying the policy makers by mail or telephone did not seem to be the way to approach this group, and our resources did not permit face-to-face interviews in different parts of the country. So we

reluctantly abandoned this approach. In our view, not including policy makers substantially underestimates CRESST influence because several CRESST researchers have worked with many different policy makers at the district, state, and national level, albeit adventitiously in most cases.

The second limitation is that we did not conduct an in-depth case study of a large unit like a state or large city school system to see how alternative assessment interacts with other factors to produce outcomes, many of which are unexpected. We could not identify a place where alternative assessment had progressed to an advanced enough degree. Ideally, a case study would examine a concrete situation in which the implementation of alternative assessment was far enough along that one could see the likelihood of ultimate success or failure. Innovations always change in the course of their implementation. The best we could do was examine the pilot projects already discussed. In omitting in-depth case study analysis, we probably have underestimated the problems that will arise in the implementation of alternative assessment.

Finally, what can one reasonably conclude about CRESST influence, impact, and effect over this five year period? First, CRESST has had a very powerful influence on two significant reference groups, researchers and test directors. Although these are not the groups who must actually do the alternative assessments, it is difficult to imagine any progress without their acceptance, support, and active participation. Without these groups, alternative assessment is unlikely to happen because these two groups exert intellectual and official authority over achievement testing. CRESST influence was achieved mostly through publications and other impersonal contacts. One must assign CRESST very high marks on impact here.

This is not to say that everyone in these significant reference groups is an advocate of alternative assessment or accepts CRESST's positions on the technical issues. A few are vociferous opponents. However, the small amount of opposition reflects how much progress has been made. Again, we note that this is alternative assessment seen as supplement, not replacement. It is also true that CRESST researchers probably find influence exercised through publications most compatible with their academic style of work. Such influence requires less adjustment on their part. The closer one moves to practitioners, however, the less publications count and the more critical personal contacts become, according to the research literature on educational change.

CRESST products have also made a measurable contribution to CRESST influence. One product, the book, has been very widely disseminated and used by practitioners, especially in teacher training. One would think that training and training materials would be critical for eventual implementation at the classroom level. It might be noted that CRESST relied heavily and successfully on a pre-established network to promote the book, thus leveraging its influence and minimizing the costs and headaches of dissemination. We take the use of an already existing practitioner network to be partly responsible for this successful dissemination.

For the second product examined, it is not clear at this point whether the CRESST model has been successfully implemented. The model is being used by some but we do not have enough reports to make an overall judgment about its success. Most opinions are highly qualified or incomplete. So on the product dimension, we would give a good but more qualified grade overall. In general, we think things become more difficult

and problems more numerous the closer one moves to implementing alternative assessments in districts and classrooms. The problem becomes one of those who must do, as opposed to those who must approve.

Finally, for teachers at the classroom level the experience has been positive, but with some sobering qualifications. Very talented researchers and developers, including four principal investigators and four graduate students, worked with fourteen teachers to develop and implement alternative assessments in one pilot project. This implementation required a substantial investment of personal time and contact on everyone's part. The development was not easy or quick. For the most part the teachers were happy afterward about the use of alternative assessment in their classrooms, even while admitting that the effort invested was very substantial on their part. The new ideas did not spread to other faculty in the participating schools.

If one looks across the country and imagines the time and effort required to implement alternative assessment in hundreds of thousands of classrooms, then the investment by teachers, researchers, developers, school districts, and states will have to be huge. One must wonder where these enormous resources will come from and how many decades such a change might take.