STATEWIDE ASSESSMENT IN CALIFORNIA

Dale Carlson

I almost feel as if I need to apologize for the distinct difference you will notice between my presentation and those of the two previous speakers we have enjoyed this morning. Whereas they have presented scholarly, theoretical papers outlining the ways in which educational measurement can lead to better instruction which in turn will lead to better human opportunities as the title of this conference suggests, my presentation will focus on a specific assessment program within one of the United States and will show in a practical way how new measurement practices and procedures have been used in developing a program.

Let me begin with an introduction to our program by showing this picture of California, rather of the United States, where the shaded section represents the State of California. In contrast to the magnitude of many of your countries, the land area and population of our country is relatively small. Furthermore, the population of California is only one-tenth of that of the United States, so I will be speaking to you about an educational evaluation system for a state which has the responsibility for educating approximately 4,500,000 children in elementary and secondary schools – children housed in approximately 4,600 schools which are operated by administrators in 1,000 school districts.

Let me then take you back to a fateful evening in October of 1961, when The Russian Space Program launched the first satellite known as Sputinik into space, event wich had the greatest impact on educational system in California and the public's perception of that educational system.

During the 1950's the United States was a place of great optimism-almost euphoria. People in the United States, although I hate to admit it, we're guilty of believing that the United States was the home of the most advanced civilization, at least technologically speaking, on the face of the earth. You can see why it came as such a such to Americans when the opposing superpower pulled the coup of launching the first satellite. Serious criticisms and investigations were launched into various aspects of our educational and scientific efforts. The public schools of the United States, and of those in California in particular, were judged to be inferior and inadequate to the needs of keeping up with the pace of the twentieth century in the world of science.

Many committees and commissions deliberated upon the problem. One recommended that the State of California develop an educational evaluation system to report annually to the people on the effectiveness of the schools as judged in part by the performance of the California children relative to those in the rest of the United States.

A system was duly ordained and implemented which continued more or less unaffected and unchanged from 1962 through 1971. That system required the assessment of children at several grade levels throughout California with a standardized achievement test in the basic skills. As all testing was done at that time, a particular test was used and all children took the very same test, the results of which were aggregated and reported to the State Board of Education and the legislative bodies. After nearly a decade of that program, criticism arose from school people that the program was neither fair nor effective. The criticisms had three specific foci. First, teachers complained that the test, being made up by test specialists in other states, was irrelevant, or at least not totally relevant in that they focused on some skills that were not being taught in California schools, or not at those grade levels. Secondly, they believed that the amount of time that was being spent on this assessment practice, that is, from three to four hours of testing per year, was an inordinate expenditure of valuable instructional time, and thirdly, in relation to the time spent in testing, the information they received as classroom teachers about their children was judged to be relatively useless.

The program which I am about to describe to you attempted to rectify those three criticisms and expand on the third: the usefulness of test information in providing a basis for improving the efctiveness of instructional programs.

A major policy decision was made in 1972 by a legislative committee appointed to study this matter. The most important decision that committee made was

to define the role of a state assessment program as one focusing on groups of children and the effects of the instructional programs on those groups, not on the impact or specific needs of individual pupils. Obviously the latter is extremely important information, but it was determined that it was not within the purview of a state agency to collect or provide that information. Local school districts would have the responsability of gathering the information appropriate to the specific goals of their instructional program while the state assessment program would focus upon those overall common goals in a more efficient manner than in the past. The greater efficiency came about through the use of matrix sampling.

Matrix sampling is a procedure for gathering information about a group of pupils in considerably less time than traditional testing practices. Since the focus of the California Assessment Program was to be on groups of students, there was no need for each student to take the very same test questions or exercises. Matrix sampling allows for the development of a long and comprehensive assessment battery, which is then stratified into a designated number of test forms, each of which has a unique set of test exercises in our application of the method This means that the testing time can be cut down to a fraction of what it would be if all students took the same test. Obviously, it would be nearly impossible for a student to take all 1,020 test items in our most recently developed test battery designed for third grade students. Each student takes on of thirty unique forms which takes only 35 minutes of his time. Each test form consists of several questions in the areas of reading, written language, and mathematics. The data, of course, do not provide individual student estimates, but that is not the focus of the program. The long test battery means that it is possible to assess a much wider array of skills and concepts than would otherwise be possible. It also has certain advantages. For example, students find it difficult to profit from looking at each other's papers, since all students within an average classroom of 30 students have a different test form! Finally, while it would still be possible for a teacher to teach the specific test items on the test to the children in order to improve the scores, it is much more difficult to do, and more importantly, much more likely to lead to real and lasting instructional benefits in the process.

This is the skeleton of the program that has been developed over the last eight years — one which uses matrix sampling to assess the content areas of reading, language, and mathematics proficiency of all children in the State of California who are in grades three, six, and twelve-end points of specific educational segments or units of instruction.

In the next section of this presentation, we will focus upon the type of information which I believe should be supplied by a comprehensive assessment program. This type of test information is, of course, just one part of what is needed to make decisions about instructional programs. We will first discuss the various types and levels of data that should be supplied, their characteristics, and then show how they happen to be supplied in the California Assessment Program. The first type I will refer to as absolute. It is perhaps only too obvious that the first information to be supplied should be some type of raw information, whether that be number of items correct, percent correct, percent of objectives mastered, or any other way that is most appropriate to the design of the measurement system. In most cases, however, the particular absolute statistic used is helpful in making longitudinal comparisons but is relatively devoid of meaning which would assist one in making a judgement about the true quality of the performance, and hence, the effectiveness of the instructional system. It is for that reason that we come to the second type of information, that of relative-general which is related to the third, that of relative-specific. In both, the process is one of drawing added meaning from an absolute score by comparing it to the scores for other schools. The difference between the two is one of norm group or comparison group. In the relative-general, the scores for a school are compared to all other schools, either in the form of a statewide average or a distribution of all schools yielding, for example, a percentile rank.

Such comparisons, for example, "How do we compare with other schools?" are almost always desired by the public and frequently by school personnel as well. However, such comparisons leave the consumer of the information desiring a more specific comparison, that is, one wants to know how the performance for a school compares to that of "similar" schools; similar, that is, in terms of the types of students that comes to that school and the resources available to the teachers. Such information can be provided in a variety of ways. The way which seems to be the most accurate is that of the development of a unique, predicted score for each school, and the placement of a standard error around that score to yield what we call a "Comparison Score Band." This approach not only has the advantage of using all of the information found in the predictors gathered such as socio-economic status, educational level of parents, and poverty index, but does not have the problems of blocking or stratifying procedures for forming groups of similar schools. It means that each school is in a sense comparing its performance with a special set of schools which in background characteristics are identical with its own. The band is arbitrarily set so that 50 percent of these hypothetical schools will fall within the band. It is, of course, a purely normative and relative operation, in contrast to a mastery or criterion referenced approach. No matter how well a set of schools is scoring, one-fourth of them will be above the comparison score band, one-fourth will be below, and one-half will be within the band. The procedure is not flawless and is, of course, certainly no better than the quality of the background information for a specific school. Nevertheles, it does serve in the minds of the public, and more importantly in the opinion of the school personnel a valuable public relations and morale function. For example, low scoring schools are not always at the bottom of any ranking, but can be shown to be doing relatively well, given the characteristics and resources available. Similarly, high scoring schools are

held accountable to a higher standard and do not receive the bouquets for superior performance merely as a function of the quality of the student body they happen to serve.

The fourth area listed on the overhead is that of trend information. Nearly all test information users agree that it is most fair and most informative to compare the performance of a school with itself. Usually the best way to do this is to compare a school score with previous performance. This means that test information must be collected frequently enough to provide a current trend line. The data must come from the last few years to gain the credibility of the users, as well as to minimize changes that may be taking place in the student body due to changes in the community.

The fifth point of information, another noncontroversial one, is that of grade levels or age levels. The value of collecting information at several grade levels is not merely one of getting a better handle on the performance of a school by gathering multiple pieces of information but has the added virtue of making it possible to chart the progress of the students as they pass through the various grade level check points. One can identify, for example, schools where the students are at a given rank or position in, the third grade, are lower at the sixth grade, and then lower at the twelfth. It would be instructive for the local faculty and staff to study their instructional program to determine the source of this pattern – to see why it is not a flat profile or why the scores, even in a relative sense, are not increasing as the students move through the grade levels.

The sixth and seventh pieces of information considered essential are scores for content areas and skill areas. Content areas are defined as major domains or objectives, such as reading, written language, and mathematics, whereas skill areas are perhaps more important in a diagnostic and improvement sense since they refer to more narrowly defined sets of skills and concepts. These subdomains of homogenously formed sets of test items focusing on specific skills are defined by teachers and other instructional personnel rather than by the preconceptions of the assessment staff. Many ways can be chosen to diplay and report the information of the students in the various skill areas. The one selected and developed in California has a relative base; it is purely an ipsative report. The school's performance on each skill area is compared to its overall performance. Skill areas are identified as relative strengths or weaknesses if they fall far above or below the total score for the content area. Therefore they are strengths or weaknesses only in a very relative sense. For example, very high scoring schools have relative weaknesses which, of course, are weaknesses in a different sense than weaknesses for low scoring schools. This is consistent with a key purpose of the assessment program: the production of information for a school which is useful to that school staff in making decisions about the effectiveness of its own program. The program obviously has an accountability and public exposure or political side to it, but we like to think that its chief virtue is that is actually leads to the improvement of the instructional programs at the local level. I will share some information with you later about this aspect.

The eighth point refers to attitudinal or affective information which has only been recently added to the California program. Other states and other countries have, of course, been collecting this for years. Suffice it to say that even if a program does not focus on such things as self-concept, respect for others, social problem-solving, etc., it can at least focus on the areas we have started to look at, that is, attitude toward reading, language, and math. The affective components of these particular skill areas are important motivating and moderating factors, in appliying the skills in problem-solving situations.

The ninth and last type of information, that of subgroup scores, is only now being experimentally reported in our program. It calls for a higher than average tolerance for a large quantity of information but it also may hold the most interesting and valuable clues about the differential effectiveness of an instructional program for those who persist to receive the fruit that can be obtained from careful and thoughtful analyses. I am refering here to the process of dividing all of the students for a grade in a school into a variety of mutually exclusive categories, for example boys and girls. In our case we also divide them into four levels of socioeconomic status, four levels of mobility, and three levels of English language fluency. Such a report would indicate to a school staff the relative performance of boys and girls mathematics versus language and how those differential patterns for a school compare to the differential patterns of other schools in the district, or to that of the statewide average. It would also yield, for example, information about the relative performance of students who have had the benefit of the school's instructional program for a number of years in contrast to the students who had just recently moved to the school – mobility of pupils being both a real and perceived problem complicating the process of determining the effectiveness of schools in a mobile state, such as California. One could also see if the instructional program had inadvertent effects in fostering relatively greater growth on the part of either higher or lower socio-economic students, inadvertent at least to all but the most sensitive observers. It is also interesting to note at this point that attitudinal information can also be treated as a subgroup variable, that is, all of the students who indicate that they do like to read or do mathematics can be reported here as a specific subgroup in contrast to those who claim to enjoy mathematics, thereby making it possible to report not only the percentage of such students but also the actual reading or math performance of those students as a group.

This then, is the end of my litany of the essential types of information that a "good" assessment program should supply.

In the United States perhaps the most topical aspect of testing at this particular point is that of the usefulness of test information. Much controversy surrounds the large quantity of testing that is taking place, and the relative lack of utility of that information in improving programs. At the local level, the state level, and the national level studies are being funded and conducted to study ways in which testing can be more easily and fre-

quently used to modify instructional programs. Therefore, a word should be said about this particular program and its usefulness. We will not focus upon the state level, although it is obvious that such a quantity of information aggregated at the state level about all of the schools and districts in the state does supply and, in fact, has been useful in typical planning roles, varying from allocating special resources to schools with large quantities of students with special needs to using the data about those schools to evaluate the impact of those very allocations.

Analysis of the statewide results have indicated specific curriculum weaknesses. For instance, in the mathematics programs across the state is was determined that the "new math" was not providing sufficient experience in computational skills. Therefore, statewide curriculum documents were revised, leading to different emphases in textbooks subsequently adopted, which in turn changed actual instructional practices.

At the local level, as I stressed earlier, use of the results is most encouraging. From a survey conducted last spring of all 4,600 elementary school principals, we were pleased to learn that over 75 percent of them indicated several analyses and uses of the information leading, we trust, to some type of school curriculum change or other type of program improvement. The overhead indicates the frequency of the various types of analyses and uses reported by the school principals. We find this particularly encouraging in light of the fact that no more than three years ago after a couple of years of presenting the reports, we had no reason to believe there was more than an infinitesimal level of use of the reports, a fact which is quite typical of test results in America. So even if the percentages shown are exaggerated by the social desirability factor, the results are still extremely encouraging, especially since they have been corroborated by various telephone interviews, personal observations and studies of specific school districts. So it is curious to note that a program which started out as a statewide accountability system relatively little statutory mention of schools and very little thought of the function of the program as supplying useful information to local school personnel, we now have a program which draws its greatest support from local personnel, especially those who have taken the time to study the results and have found them useful in improving their local programs.