SURVEY ON ECIA CHAPTER 1 EVALUATION REGULATIONS

CSE Technical Report 293

Sharon Johnson-Lewis, Chairperson

CRESST Chapter 1 Evaluation
Regulations Study Group

UCLA Center for Research on Evaluation,
Standards, and Student Testing

July, 1988

CRESST Chapter 1 Evaluation Regulations Study Group

Committee Members

Sharon Johnson-Lewis, Chairperson
Detroit Public Schools

Gary Thompson, Columbus Public Schools

G. Kasten Tallmadge, RMC Research Corporation

Sandra Pakes, Psychological Corporation

James B. Olsen, World Institute for Computer Assisted Training

Carl Novak, Lincoln Public Schools

Robert Nearine, Hartford Public Schools

Steven Murray, Northwest Regional Educational Laboratory

Roger Mitchell, National Urban League

Robert Linn, University of Colorado

Daniel Levine, University of Missouri

Nancy Enell Law, Sacramento Public Schools

Ofelia Halasa, Cleveland Public Schools

Stephen H. Davidoff, Philadelphia Public Schools

Report on the CRESST Survey
Chapter 1 Evaluation Regulations Recommendations

Introduction

In June, 1988, the Chapter 1 Evaluation Regulations Study Group of the Center for Research on Evaluation, Standards, and Student Testing (CRESST) distributed a survey to local and state education agencies, universities, and Technical Assistance Centers throughout the United States.

The survey was intended to give respondents an opportunity to react constructively to the evaluation requirements of the Education Consolidation and Improvement Act (ECIA), Chapter 1. The Evaluation Regulations Recommendations Survey was composed of two parts. Part I contained fourteen questions followed by either suggested responses and opportunities to comment or no suggested responses, only opportunities to write comments. Part II of the survey listed key Chapter 1 evaluation issues that were to act as idea generators for position papers.

One hundred eighty surveys were distributed and thirty-four responses to Part I of the survey were received. Part II of the survey generated one position paper. In addition, position papers were solicited by personal invitation from members of the educational research and evaluation community. This latter process generated six more papers for a total of seven.

Presented on the following pages are the fourteen questions contained in Part I of the survey, the respondents' choices to suggested responses, and the respondents' comments. Following the presentation of responses there is a set of recommendations based on the survey. The final section of this report consists of the position papers, each of which contains its own set of recommendations.

Presentation of Responses

Presentation of Responses

The first item on the survey inquired,

> To what extent is Chapter 1 data, as currently required, useful for program monitoring and improvement?

Respondents rated its usefulness on a scale ranging from 1 to 5 in which the numeral 3 represented average usefulness. Of the 33 respondents, 7 gave a rating of 2, 12 respondents rated the usefulness as average, 9 gave a rating of 4, and 5 rated the required Chapter 1 data as very useful. The overall rating was 3.4. This item was the only one in the survey using a Likert scale.

Comments from the respondents provide insights about the issue of the usefulness of mandated Chapter 1 data for program monitoring and improvement. In general, most respondents felt that the Chapter 1 data which are required for state and for federal reporting serve those purposes well. However, they also felt that the attributes of such data in isolation from other kinds of information limited their appropriateness for program monitoring and program improvement. The following remarks were sampled from the comments addressing this issue.

> Chapter 1 data as currently required is only moderately useful. Our local evaluations include much more detailed reports which reflect school level data. The reports offer information used for program improvement, but are generally used by central administration and/or decision making committees. The data does not have much value for monitoring other than on a yearly basis.
>
> We report on our Chapter 1 achievement gains more extensively than required and in ways that are more suited to our district needs.
>
> While data collected is useful, use of the data for program monitoring is minimal.
>
> Required evaluation does not provide timely, on-going data for program monitoring. Because of high student mobility, even pre- post data on an annual basis provides only limited data for program improvement.
>
> Current data requirements, although limited in terms of usefulness for monitoring and improvement, should not be entirely dismissed. Annual evaluation results should certainly be used as one of several measures to identify unsuccessful programs. Other criteria such as skill mastery, graduation rates, dropout rates, performance levels, etc., could be used in addition to test scores.

The second item on the survey asked the question,

> What suggestion would you have for an alternative evaluation system to the Title I Evaluation and Reporting System (TIERS), keeping in mind data must be aggregated nationally.

In general, respondents did not want the current Title I Evaluation and Reporting System replaced. They felt that it had attained status and that it is a system with which school districts are now knowledgeable. In addition, over the years a great deal of time and money had been invested on inservice training related to the implementation of TIERS and the use of results. Many also felt that it is a sound approach to the aggregation of data nationally. A sampling of comments from the respondents follows.

> Any new models would require extensive time for development and piloting before they would be deemed appropriate for national aggregation. Although other systems could be developed, they would need to be substantially better then TIERS to warrant the training of all SEA's and LEA's in implementing the model.

> Keep the TIERS system, but rename it CHIERS for Chapter 1 Evaluation and Reporting System. This is a system that our school districts are familiar with, and after spending a lot of time and dollars on inservicing them in correct implementation and use of results for program improvement, they now implement and use it correctly. Also, the results can be aggregated nationally.

> Why change a system that works! The TIERS model allows local decisions concerning which tests to use, yet allows for national aggregation.

> NCE gains are not particularly useful locally, but do serve the purpose for federal accountability.

> If it is necessary to report all data for national aggregation, TIERS appears to be working.

The third issue in the survey was related to the Gap Reduction Model as an alternative to the Title I Evaluation and Reporting System. As written the item read,

> Gap Reduction Model calculations have been recommended as an alternative to the TIERS Model A normal curve equivalent gain calculation. Would you be willing to participate on a trial basis for one or two years to report Chapter 1 evaluation results using both models?

Thirteen of the 34 respondents indicated a willingness to participate on a trial basis; however, most had reservations. These reservations centered around anticipation of additional burdens upon schools as well as indications of a lack of knowledge of the Gap Reduction Model. Fifteen persons responded that they wanted to do the TIERS model calculations only. Of the 6 remaining respondents, 3 had concerns about gap reduction calculations and the other 3 requested additional information or technical assistance. Comments accompanying this issue were,

> The Northwest Regional Educational Laboratory has been undertaking a study in cooperation with the Washington SEA. It has found that the Gap Reduction results (a)are not highly correlated with NCE gains, (b)depend greatly on which test is used, and (c)vary greatly from grade to grade when the relative growth index is used as the outcome metric.
>
> As a TAC person with a general knowledge of the Gap Reduction Model, I would have definite concerns about implementing the model for national aggregation purposes. Although the model sounds interesting, there are also definite concerns that I have about how the gap reduction would work when different norm-referenced tests are used. Some evaluators when applying actual data to the model have found considerable disparity with the numbers when different NRT results are plugged into the formula. I think the model has considerable utility at the local level, especially for early childhood and certain types of migrant evaluations.
>
> We have done lots of gap reduction calculations. What we learned is that how you calculate a relative growth index (RGI) depends on what reference you use. We know of at least 3 different calculations for RGI. We have also learned that RGI's are difficult to interpret. My intuition suggests the RGI is not equal interval either. Besides, when you use the national norms for a control group, the gap reduction model is the same as the TIERS Model A. Conceptually, the model is nice, but we have too little experience to use it beyond a simple tryout.
>
> A comparison of relative growth indices from the same percentiles but using different tests reveals a wide variation, a fact which certainly negates their use for Chapter 1 national reporting purposes. Another pertinent question is, "Does the Gap Reduction Model stimulate any better treatment or remedy for Chapter 1 programs than TIERS?"
>
> Implementation of the Gap Reduction Model should not be considered until the model is proven to provide more information than the TIERS model. Each time procedures

change, errors abound. The consistent use of TIERS means that the quality of the data is better. Training sessions would be imperative if any changes take place.

I am not opposed to studying other models, but this state has reported on the TIERS model since its beginning. All LEA's are using the TIERS at this time. Why change a system that works! The Gap Reduction Model would require more training. Besides, aggregation from one test to another has a major error.

Item 4 addressed the issue of whether criterion-referenced tests could serve a dual purpose, (1)evaluation of programs, and (2)meet the requirements of providing test data which could be aggregated to the national level. The question as stated in the survey was,

> How might the use of criterion-referenced tests for evaluation purposes meet the requirements for aggregation of data to the national level?

Sixteen of the respondents said it should not be attempted or that it could not be done, 11 suggested linking the criterion-referenced test to a norm-referenced test via an equating procedure, and the rest were divided between miscellaneous suggestions and no choice at all. Some of the pertinent comments were,

> CRT's are good for evaluating instruction, deciding whether students learned as a result of instruction. Nonetheless, the goal of all Chapter 1 programs is to effect achievement in a more global fashion. NRT's provide a broader, and more useful, sampling of content for evaluation.
>
> Very few districts will have the capability to link local CRT's to NRT's. Very few SEA's will have this capability. This position comes from the fact that our TAC has observed the problems that test publishers have in linking one edition of a test with another.
>
> Test equating is not possible for the great majority of districts due to small sample size. When medium to large districts have undertaken equating, the results have often been perplexing.
>
> Equating should be used only if it can be clearly demonstrated that accuracy obtained is within acceptable limits.

Item 5 of the survey read,

How can the progress of students in the area of "more advanced skills" (e.g., reasoning, analysis, interpretation, problem solving, and decision making) best be measured, assuming results must be aggregated to the national level?

Comments accompanying responses to this issue can be placed in three categories. First, some respondents felt that Chapter 1 was designed to teach basic skills to those students who needed remediation in reading and mathematics. Therefore, the teaching of more advanced skills would be inappropriate. A second thought was that higher level skills should be part of the content areas taught and that measurement of these skills is embedded in norm-referenced tests currently in use. The third idea is closely related to the second. Even if more advanced skills are emphasized, there are no adequate tests on the market for their measurement. In many cases, comments from a respondent encompassed more than one category.

There is a need for development of a higher order thinking skills test rather than just a classification ordering of existing standardized test items.

These skills are embedded in current tests to some extent. There are no good thinking skills tests on the market. Besides, higher level skills should be part of content areas taught.

If a national requirement exists, then NRT tests could be used. However, it would require the entire test battery, science, social studies, etc., to be administered because advanced skills are typically embedded in all of the subtests. If this notion becomes a reporting requirement, NRT's are the only option available for national aggregation. However, I would not recommend attempting the national aggregation of test scores of items which measure advanced skills.

This is a problem for test publishers. Advanced skills are embedded within the total test battery. Chapter 1 instructional programs could be designed to provide opportunities for children to experience higher order thinking skills.

Require that the comprehension subtest on standardized tests be given and the score for comprehension be reported nationally. Ditto for problem solving in mathematics. This way, little additional testing would be necessary and these 2 more advanced skills could be reported and aggregated nationally. This would be preferable to requiring that an entire test battery be administered and the higher order thinking items be flagged and a "HOTS" score derived. Such a requirement would put an additional burden on schools.

> The focus of Chapter 1 is remedial. Limit evaluation to the measurement of basic achievement skills.
>
> Wrong question! Chapter 1 is designed to teach basic skills in reading and mathematics. Leave the more advanced skills to more advanced classes.
>
> The Chapter 1 students served in our schools score at or below the 25th percentile and are generally 2-3 years below grade level by the 6th and 7th grades. Nearly 20% of our Chapter 1 students continue more than one year, so basically we have a new group of students each year. Granted all students need to have exposure to the advanced skills, but I am not sure we systematically teach them.

The aim of Item 6 was directed at obtaining feedback on the impact on local testing programs of a national regulation requiring Spring to Spring or Fall to Fall testing to measure the achievement of Chapter 1 participants. The Item was worded as follows:

> What impact would a national standard requiring student achievement to be derived from annual testing (Spring/Spring or Fall/Fall) have on your local testing program?

Most respondents favored Spring to Spring testing, but many of these and others acknowledged the need for a Fall to Spring cycle as well. In school districts having high transiency, the loss of students over the summer and the influx of others would necessitate a Fall testing program in order to have an adequate number of premeasures on students being served by Chapter 1. It is also not uncommon for school districts to have a regular testing program every Spring for only a few grades. That is, not every grade level is tested every Spring; hence, it is necessary for them to conduct testing in the Fall. A few respondents indicated that Fall to Spring scores can be adjusted to yield results which would be comparable to Spring to Spring scores. Fall to Fall testing did not receive support in this survey. Only 2 respondents mentioned it and then only to say that this cycle was not used.

Comments from the respondents relative to this issue follow.

> Some schools have a relatively high transiency rate for students and families, and obtaining an "N" of sufficient size under a mandated annual testing could be difficult or impossible.
>
> Requirement of Spring/Spring could have a great impact on LEA's by requiring an additional test in 6 grades for Chapter 1 students.

> We are a very transient area and need a Fall/Spring testing cycle.
>
> In the last 3 years there has been steady movement in the direction of an annual test cycle. Those holding on to the Fall/Spring model give population transiency as the reason. Requiring annual assessment would force their hand.
>
> From a TAC standpoint it should not create any additional testing or reporting burden on SEA's or LEA's. Even if districts choose to evaluate on a Fall/Spring (which under certain circumstances may the most appropriate cycle) reporting Spring/Spring results could still be done with little effort. Just use the Spring results to conduct the evaluation. Formulas have also been developed that allow Fall results to be used to predict the previous Spring results.
>
> While annual testing is the norm in this State and ought to be mandated nationally, there will always be some exceptions. For example, districts sometimes have gaps in their testing program that make Fall/Spring evaluations necessary. In these cases, results can be adjusted to yield results which are comparable to those obtained via Spring/Spring.
>
> It is more practical for some LEA's to implement a Fall/Spring testing cycle because of high transiency rates and coordination with local testing schedules. This flexibility for data reporting ought to be maintained.

Item 7 posed the question,

> What viable alternative is there to annual reporting, keeping in mind evaluation results should constantly be used for local program improvement?

Thirty of the 34 respondents felt that there was no viable alternative to annual reporting, and/or they endorsed annual reporting of evaluation results. The other 4 respondents suggested either an every other year or a 3-year cycle with the "off" years being used to emphasize process evaluations. Comments included the following:

> We see no viable alternatives to annual reporting if systematic efforts at program improvement are to be successful.
>
> Our state endorses annual reporting of evaluation results. The uniformity of statewide testing and implementation of computer technology facilitate

annual reporting.

I don't think there is an alternative. Most people seem to work to improve the quality of their current program. They do not want to think about change.

Since the new authorization states that progress must be monitored annually for program improvement, I do not see any way to avoid annual reporting.

The TIERS Model A results appear to be the best way for LEA's to conduct annual evaluations to meet the program improvement requirements of the law. If LEA's evaluate annually, reporting the data annually should be no additional burden. Whether or not annual reporting is required, our State rule making will probably require annual reporting.

The subject of sampling was approached in Item 8.

In what respects would sampling procedures be adequate for reporting or for use in local program improvement efforts.

Overwhelmingly, the survey respondents felt that sampling would not provide sufficient information. Reasons cited were

1. The need to assess individual student progress precludes sampling procedures for testing,

2. Program evaluation at the school site becomes low quality if all Chapter 1 participants are not included, and

3. In most cases the counts of participants are small. Sampling would not only reduce these counts further, but also reduce local confidence and ownership of results.

Five respondents indicated some support for sampling procedures. However, the majority of these qualified their responses by pointing out that the procedure might be adequate for total system reporting or that it might be appropriate in large school districts. Some representative quotes which have been extracted from the comments follow.

Sampling never works because ultimately you want to use test scores at the individual level. Under the new Chapter 1 law, districts must have individual program improvement plans if they are not making gains in the program. How could you possibly do this without periodic testing. If you are testing periodically, you might as well report annually. If you remove evaluation

for some students or schools, you destroy routine. Then when it comes time to do an evaluation, it turns out to be low quality. I'm a strong believer in sampling too! It just doesn't work in Chapter 1.

For reporting purposes sampling procedures could be adequate, even for aggregation to the national level. Local program improvement efforts would undoubtedly be questionable in terms of appropriate practices, especially in view of identifying those individual students for whom Chapter 1 does not appear to be working. For improving weak programs, sampling is not appropriate.

Due to the very high mobility in the Chapter 1 population, only about 65% of participating children are currently reflected in evaluation reporting. Sampling would (a)reduce this number while reducing local confidence and ownership of results, and (b)produce more rather than less work.

For local program improvement, sampling procedures are quite adequate to provide information about sustained gains. Guidelines to collect adequate data should be established to ensure analysis and sound interpretation of the results. If the objective is local program improvement, then the flexibility of the LEA to design a meaningful study must be maintained. Therefore, reporting using sampling procedures becomes less of an issue.

The numbers of students in our State are not massive, so we would not benefit from sampling and in some cases would not get enough information. Also, the requirement to assess individual student progress militates against sampling.

The ninth Item on the survey inquired

> In what ways are Chapter 1 demographic data (gender, age, ethnicity), which is currently required, useful information for local program improvement?

Eleven respondents thought that such information was useful for disaggregation of data by group and that such disaggregation would allow for identification of strengths and weaknesses of the program by group. Twenty-one respondents suggested that the demographic data were not useful or at most of limited use. Their reasons included the heavy burden on LEA's for data collection and for analysis. Another reason cited was the low numbers of minority groups within a State. The following comments were made relative to this issue.

The disaggregation of data by groups allows for identification of strengths and weaknesses of the program by group.

Currently of very limited use. The only way to make it more useful would be crosstabulations with other data elements. However, that would create a tremendous data collection burden for LEA's and that is the trade-off. More data collection for more data uses. Is the trade-off worth it? Probably the answer is yes, but considerable thought would need to be given to the delicate balance between data collection and data usefulness.

Gender and racial/ethnic data are useful for group comparisons; age data as currently defined (year of birth) are worthless.

Useful information to have even though we are not required to cross tabulate it (thank goodness). Helps in determining where to focus resources.

Not good data in a low minority State like ours.

Until all systems have computer programs that will aggregate data for selected groups, this data will not be used for program improvement.

At the district level the usefulness of this information is limited to reporting purposes only. Use of data for program improvement is practically nil.

The most useful demographic data is on ethnicity since we monitor achievement for each group. Age and gender are rarely used for local program improvement.

Our district does not and will not disaggregate achievement data by race/ethnicity.

In small districts the usefulness of such data may be minimal.

Item 10 read

How could teacher ratings, grades, classroom tests, and performance inventories be used in Chapter 1 evaluations?

Thirty-two of the respondents mentioned uses for these measures in Chapter 1 evaluations. Some of these applications were

1. Supporting information for the regular evaluation,

2. As other measures to determine program effectiveness,

3. As local criteria to assist in program improvement and changes for individual kids,

4. To expand interpretation of achievement scores,

5. To support/reinforce standardized measures,

6. As measures of progress on specific objectives of a project,

7. As measures in determining student's eligibility for for the program, and

8. As feedback to the teacher during the school year.

Two respondents felt that the measures listed in the survey question should not be used. Their comments speak for them.

> Data collection and reporting burden would not justify the information gathered.

> To use the instruments listed above would just muddy up Chapter 1 evaluations. The subjective nature of those instruments result in biased results.

Item 11 posed the question,

> What impacts do Chapter 1 evaluation requirements have on the local testing program?

Responses to this issue reveal a great deal of variance in the way school districts cope with the requirement. The problem is exasperated by several factors. Very few school districts have annual testing programs which test every student at every grade level. Testing programs are generally unique. Some districts administer norm-referenced tests only and some may administer a mixture of norm-referenced and criterion-referenced tests.

Additionally, some States mandate their own testing programs which are usually criterion-referenced. Superimposed are Chapter 1 requirements, testing for selection, pretesting, and posttesting.

From the responses it appears that school districts report results from their regular testing programs whenever possible. The gaps are then filled by additional testing in order to fulfill requirements. The following quotes are representative of comments made by respondents.

> There is a need to consolidate and integrate norm-referenced testing, criterion-referenced testing, and federal categorical program testing requirements.

> Most districts in the western U.S. use their district testing plus an additional Chapter 1 administration of the same test. Children are thus tested twice as often in many districts, but the effect is not to duplicate testing or to create a parallel system.
>
> Spring/Spring evaluations usually capitalize on existing testing programs in school districts, thereby reducing testing burden associated with Fall/Spring cycles.
>
> Chapter 1 requirements have required some districts to use more than one test or to double test. Most districts have used Chapter 1 requirements to improve and simplify testing procedures.
>
> In our State we work toward coordination of Chapter 1 evaluation requirements with local testing, i.e., Chapter 1 programs use local testing program results to report.
>
> With high school proficiency exams, minimum competency tests, State mandated achievement/ability testing, district-wide annual achievement testing, Chapter 1 selection testing, Chapter 1 pretesting, Chapter 1 posttesting, NAEP, participation in norming studies, in-house research, national research, local university research, special education testing, gifted and talented testing, etc., etc., etc., anything that can be done to reduce testing should be done.

Item 12 was intended to obtain ideas on evaluation requirements for early childhood programs. The question was worded,

> What evaluation requirements, if any, should there be for early childhood programs?

Six respondents felt that the evaluation requirements should parallel other Chapter 1 requirements. Fourteen respondents indicated that demographic data only should be required, and 11 others had a mixture of suggestions. Demographic data was felt to be not only easily aggregable at the State and national levels but also descriptive of the population being served. Suggestions of other approaches to evaluations included

1. evaluations based on local objectives,
2. longitudinal studies,
3. evaluations based on developmental skills, and
4. evaluations of selection procedures.

Respondents in the latter category made it clear that aggregation should not be required. Comments from respondents on this issue follow.

Aggregation to the State and national levels should not be required. Evaluation on an annual basis should be required in view of the growing number of dollars spent on these programs. Sustained effects studies should not be required. More emphasis should be placed on selection procedures as many more students are being served than really should be served.

Longitudinal follow through into the 2nd grade.

Early childhood programs should be evaluated in ways and using instruments that meet local needs. I know of no standardized test for young children that is valid and reliable and aggregable. Locally developed checklists of skills and performance accomplishments can assess young children's achievement and development. For national purposes I'd recommend conducting a series of case studies, naturalistic in form. Let's not subject young children to paper and pencil tests which serve to disturb and frustrate.

Each district should develop the appropriate procedures to evaluate the program according to objectives selected.

The requirements for preschool and grade 1 should not parallel other Chapter 1 requirements because the data is unreliable. We should not encourage inappropriate practices. Because the law is clear that achievement data is not required for preschool and grade 1, the requirements for these groups should be left to State and local decision. The collection of demographic data on these participants is not difficult to obtain or aggregate and should be collected to describe the participants.

There should not be any exceptions made for Chapter 1 requirements.

Districts need to do entry level diagnostic testing that can be related over time to students failing kindergarten, going into special education, requiring Chapter 1 remediation in primary grades, etc., so that an evaluation can tell first of all whether the program is identifying and serving the proper population. Achievement test data and achievement growth are useless at this age. The tests are too unreliable.

Item 13 asked the question,

What evaluation requirements, if any, should there be for migrant programs?

Very few of the respondents were involved with migrant programs or evaluation of migrant programs. However, four of the respondents to this question had consensus in their recommendations. Their suggestion was to develop a national evaluation in which the same standardized test would be administered at the same time by all school districts which had migrant programs. Demographics would be reported for the same period, and all data would be fed into a national database for analysis.

One of the more thorough explanations submitted by a respondent is being submitted as a position paper rather than as a response to the survey.

The final question in the survey was,

> In what ways should procedures used to measure sustained effects be modified?

Only one-half of the 34 respondents to the survey responded to this issue. Responses ranged from suggestions that the sustained effects studies be discontinued to making it an annual requirement to keeping it on a 3-year cycle. There was no strong feeling of consensus for the trend of thinking. However, there were a few compelling comments.

> Sustained effects requirements should require <u>only</u> comparing the inclass performance of exited Chapter 1 students with that of their peers. No 3 data points. No tests. The real issue, at least the one that resulted in the sustained effects requirement, is the performance of students after leaving Chapter 1.
>
> Districts try to ignore the requirement because they have not institutionalized procedures to do it routinely. By requiring it annually, they will develop procedures with improved capacity to undertake other longitudinal studies as well.
>
> Should be conducted once every three years. We do not need to do a sustained effect study every year.
>
> It is this respondent's belief that the Sustained Effects Studies (SES) are, in the majority of cases, nothing more than a paper process. The U.S. Department of Education does not collect and use the data, the SEA does not collect and use the data, most LEA's, lacking the expertise to conduct or understand the SES, do not use the data except to be in compliance. It is believed that the U.S. Department of Education and TAC's lack understanding of HOW TO INTERPRET AND USE the SES and consequently so do SEA and LEA personnel. The SES requirement should be eliminated altogether unless the

process can be taught and the information used without excessive cost and additional paperwork.

Recommendations

Based upon the responses and comments to the CRESST Chapter 1 Evaluation Regulations Survey, the following recommendations are being made.

1. There should be no change or modification to the Chapter 1 data requirements with the exception that the demographic data, which a majority of the respondents suggested was not useful or was of limited use, should be deleted. Annual evaluation results for all participants should be required.

2. The Title I Evaluation and Reporting System (TIERS) should not be replaced by an alternative system, and the Gap Reduction Model should not replace TIERS.

3. Norm-referenced test results should continue to be used for the aggregation of data at the national level. Whereas, local education agencies may choose to utilize criterion-referenced tests, teacher ratings, grades, classroom tests, and performance inventories to supplement their local evaluations, these measurements should not be considered for evaluation at the national level.

4. If the measurement of "more advanced skills" were to become a requirement, then items already embedded by most publishers in norm-referenced tests should be utilized.

5. A national standard that student achievement be derived from annual testing (Spring/Spring or Fall/Fall) should not be required.

6. Early childhood programs should not be evaluated using achievement tests. Demographic data only should be reported for aggregation at the national level.

7. There should be a national evaluation plan for migrant programs. The plan should include a common test administered to all participants in migrant programs in the Spring of each year.

Position Papers

Position Papers

Table of Contents

Use of Norm-Referenced Tests to Measure Higher Order Thinking Skills

Dan Levine
University of Missouri

Several lines of research have converged to indicate that Chapter 1 instruction frequently overemphasizes basic mechanical skills and neglects development of thinking and other higher order skills among many disadvantaged low achieving students. (Allington, 1988). For example, the National Assessment of Chapter 1 (Birman, et al, 1987) found that "Chapter 1 projects provide students with few opportunities to engage in higher order academic skills" (p. 8), and a major Chapter 1 study conducted at the Far West Laboratory (Rowan, et al, 1986) reached the same conclusion and recommended that Chapter 1 projects at all levels should expose students to higher order thinking skills, especially opportunities to read connected text and to apply mathematics to real world problems" (p. 9.7).

Problems involved in testing higher-order skills in Chapter 1 are important in evaluation because lack of adequate tests can reinforce unproductive tendencies to overemphasize mechanical skills. Among the major goals and functions of Chapter 1 evaluation are to: 1) help identify and select students who should participate in and can benefit from Chapter 1 services (Davis, 1987); 2) help assess gains in learning and identify skills for which additional improvement is needed; and 3) thereby serve as a stimulus in re-directing instruction toward greater emphasis on higher order skills. Lack of adequate tests of higher order skills obviously impedes identification of students who need assistance in this area, multiplies the probability that sites with unproductive emphasis on lower-order skills will be assessed as successful, and fails to provide appropriate information for re-directing instruction toward more stress on thinking.

Both norm-referenced tests designed to identify a student's or groups's performance level in comparison with other students or groups and criterion-referenced tests designed to determine whether a student or group has mastered a given criterion skill or set of skills have advantages and disadvantages as regards Chapter 1 testing of higher-order skills, as is of course true in general of other testing purposes and functions. For example, norm-referenced tests generally are of relatively little use in determining how far below the standard low-achieving students perform, and what skills are most important in moving them toward the standard. In many cases about all one can say with confidence is that the student or group could not read either the pre-test or the post-test or both, thus making it difficult or impossible to determine whether they made meaningful gains or whether some

students require more or different interventions than others. Furthermore, students who cannot read the test are likely to be viewed as lacking basic mechanical skills even though they may have sufficient mastery of those skills to justify re-direction toward greater emphasis on thinking and other higher order skills.

Criterion reference tests, on the other hand, also tend to reinforce overemphasis on lower order skills unless the criterion skills for which performance is assessed deal with thinking and other higher order skills. If all or most of what is assessed are lower order skills, test results may identify many deficiencies for students who did poorly in part because they not succeeded in constructing a meaningful framework within which to process and retain information and understandings (i.e. higher-order processing skills); in any case, such tests hardly can identify higher-order skill deficiencies which are not part of their content and concern.

Noting the deficiencies associated with both norm- and criterion-referenced tests in relation to assessment of higher-order skills of Chapter 1 students, evaluation personnel should support and help in the development and utilization of tests that provide data regarding both normative performance and criterion-mastery of higher-order skills. Norm-referenced data can help determine whether a student or group is succeeding in meeting standards attained elsewhere by other students, and criterion-referenced data can help identify students' functional performance levels and deficiencies with respect to higher order skills. Fortunately, tests which address both aspects are in various states of development, and proposals for greatly expanding such efforts are beginning to receive national attention (Alexander and James, 1980).

For example, the National Assessment of Education Progress has categorized performance levels in its reading and math tests so that they can serve both norm-referenced and criterion-reference functions. (The performance levels distinguish between low-order, "rudimentary" skills and higher order skills involving processing and application of information in reading and math.) Similarly, the College Board has developed the Degrees of Reading Power Test, which provides standardized scores portraying individual and group performance in comparison to national norms and also provides scores indicating functional level of mastery in comprehending test, and test developers in Illinois, Michigan, and elsewhere are making progress in constructing tests that may overcome some of the deficiencies currently associated with norm- and criterion-referenced testing. (Linn, 1988). In addition, even tests which are now most widely used can be utilized in a more sophisticated and effective manner in assessing higher-order skill development in order to improve the operation of

Chapter 1. Accordingly, the following recommendations and suggestions are offered with respect to testing of higher order skills in Chapter 1.

Recommendation/Suggestions:

1. Chapter 1 personnel and projects should support and participate in the development of improved tests that will help overcome the deficiencies of current tests with respect to assessing higher-order skills of low achieving students.

2. Steps should be taken to ensure that Chapter 1 students can read the tests that are administered to them, partly in order to help determine whether their learning problems center on non-acquisition of basic mechanical skills or on poor performance with respect to thinking, comprehension, and other higher-order skills. In addition to introduction of improved tests recently developed or now being developed, steps taken might include more out-of-level testing and greater utilization of and emphasis on sub-scores (e.g. reading comprehension distinguished from language mechanics).

3. If current norm-referenced tests continue in use in Chapter 1 programs criterion-referenced tests that assess students' functional level with respect to higher order tests also should be used. One example of such an instrument is the Degrees of Reading Power test that assesses students' skills in processing text at a defined level of competence. Among other purposes, such tests should be used to make sure that students performing fairly well with respect to higher order skills are not assigned to instructional settings that overemphasize low order skills to "solve" a non-existent problem.

4. Students' performance with respect to tests of thinking and other higher order skills should play a part in selecting those who receive Chapter 1 services, as well as the nature of the services they receive.

References

Alexander, L. and James, H.T. (1987) The Nation's Report Card: Improving the Assessment of Student Achievement. Cambridge, MA.: National Academy of Education.

Allington, R.L. (1988). How Policy and Regulation Influence Instruction for At-Risk Learners or Why Poor Readers Rarely Comprehend Well and Probably Never Will. In B.F. Jones and L. Idol (Eds.), Dimension of Thinking and Cognitive Instruction. Hillsdale, N.J.: Erlbaum.

Birman, B.F., et al. (1987). The Current Operation of the Chapter 1 Program: Final Report from the National Assessment of Chapter 1. Washington, D.C.: U.S. Government Printing Office.

Davis, A. (1987). A Meta-Evaluation of Chapter 1 Student Selection. Paper presented at the annual meeting of the American Educational Research Association, April 24.

Linn, R.L. (1988). Dimension of Thinking: Implications for Testing. In B.F. Jones and L. Idol (Eds.), Dimension of Thinking and Cognitive Instruction. Hillsdale, N.J.: Erlbaum.

Rowan, B., Guthrie, L.F., Lee, G.V., and Guthrie, G.P. (1986). The Design and Implementation of Chapter 1 Instructional Services: A Study of 24 Schools. San Francisco: Far West Laboratory for Educational Research and Development.

Sources of Errors in the Chapter 1 Reporting System

A Position Paper Prepared for the Chapter 1 Evaluation Regulations Study Group of the National Faculty of the Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Prepared by

Gary Thompson, Ph.D.
Columbus Public Schools
Department of Evaluation Services
52 Starling Street
Columbus, Ohio 43215

Introduction

The purposes of this paper are to review the Chapter 1 reporting forms in a state, to suggest sources of errors by LEAs in completing the forms, to suggest sources of error in compiling data across LEAs, and to suggest ways of reducing errors in the Chapter 1 data. The forms used by the Chapter 1 office in Ohio were reviewed because the author is familiar with the forms and had immediate access to them. It is assumed that the forms vary from state to state. However, some of the data fields must be in common across SEAs to permit aggregation of Chapter 1 data beyond the state level.

In the remainder of this paper, the author attempts to identify sources of error in the data and/or in the use of the data at the SEA and LEA levels. The sources of error described in this paper are a combination of actual errors observed by the author and scenarios contrived by the author. None of the situations described are intended as criticism of specific districts, agencies, or individuals.

LEA Sources of Error

In this section, four sources of error are identified and discussed. Figure 1 is a summary of the sources of error. In the remainder of this section, each source is described, and procedures are suggested for reducing the errors from each source.

Source of Error
° Improper scoring of tests
° Improper reporting of scores
° Lack of attention to test fit
° Lack of formal evaluation training

Figure 1

Summary Description of Sources of Error in Chapter 1 Report Originating Within the LEA

In Ohio, there are over 600 LEAs of varying size. Only a few larger LEAs have evaluation departments with personnel trained in evaluation, testing, and data analysis. In most LEAs, Chapter 1 evaluation is one of many duties of a certificated staff member who has little formal evaluation training. Even under the best of conditions, of time and job pressures, these people have difficulty completing the Chapter 1 reports. They frequently do not have computers available to help them compile data, score tests, etc. The combination of their lack of training with the absence of appropriate equipment to reduce the potential for computation errors is a significant source of error in Chapter 1 data collection.

The author's position is that the SEA should share with the LEA the the responsibility of training the evaluators. Annual workshops on the Chapter 1 reports should be funded by the SEA along with opportunities to attend workshops on other evaluation topics for a nominal fee to the LEA.

Guidelines for Chapter 1 evaluation include guidelines for determining the degree to which a NRT fits the population of students tested. The guidelines involve the simple determination of the percentage of students scoring at or below the guess level or above the ceiling level. As a rule of thumb, if more than ten percent of the students score at or below the guess level, the test is too difficult for the population. Although Chapter 1 reporting does not require reporting of NRT scores below grade 2, in Ohio, many districts have a Reading Recovery Program in grade 1 funded from Chapter 1, and there is a desire to collect data on that program across the state. However,

measurement of reading among low achieving first grade students is, at best, difficult. The author has studied various combinations of pretests and posttest with first grade students without finding a solution that is satisfactory to all parties involved in Reading Recovery. Reading Recovery program personnel want the pretest and posttest to include tests of reading comprehension. However, in September, low achieving first grade students cannot read. In the author's experience, a very large percentage (42-70%) score at or below the guess level on a test of reading comprehension. That's not an appropriate measure for a pretest. Again, in the author's experience, a test of reading comprehension has a reasonable fit at the end of the first grade. Some LEAs do report pretest-posttest gains on reading comprehension, and based on the author's experience, the resulting gains are spuriously high.

A part of the reporting procedure for Chapter 1 should include a reporting of the degree of fit of the tests used. Data for which the degree of fit is inappropriate should be identified as not suitable for aggregation.

The scenario for the improper scoring of tests as a source of error is related to the scenario just portrayed for lack of attention to test fit. The manuals for many NRTs include explicit instructions for determining if a student's test is invalid. For instance, instructions for determining if a student's test is valid may involve the number of items attempted and the number of items correct. However, if the instructions are ignored, data which should be held as missing are included in reports. The resulting NCE scores are probably very low.

If the low scores are pretest scores and the students posttest is valid, the resulting gain scores will be spuriously high.

The scores reported and analyzed by the LEA are the basis for all aggregation of Chapter 1 achievement data. If the LEA reports inaccurate data, there is no opportunity for the aggregated data to be accurate. The earlier discussion regarding lack of attention to test fit and improper scoring of tests are obvious sources of inaccurate data. However, another scenario involves an evaluator who, recognizing that the pretest score are spuriously low, improperly corrects by recording an NCE score of one or zero for the pretest, records the posttest, and computes a spuriously high gain score. In the author's state this error could be detected by the SEA in the years in which the LEA must report student by student data. However, in non-reporting years the error could not be detected. Likewise, the error will not be detected if edits are not applied to the LEA student by student reports.

The reduction of this source of error in LEA reports could be significantly facilitated if LEAs were required to report data in a processible format. The author suggests that a limited number of formats could be established for submitting data in processible format. For example, IBM or ANSI tapes are commonly exchanged across mainframe computers, and standard exchange formats for microcomputers such as DIF format or ASCII text files are becoming common among software packages. The edits suggested by the author could be effected using paper reporting systems but would be substantially enhanced by delivery of data in a processible format.

SEA Sources of Error

In this section, seven sources of error are identified and discussed. Figure 2 is a summary of the sources of error.

Preparation of a set of clear directions that will be accurately interpreted by over 600 persons with widely varying levels of evaluation knowledge and skills is a challenge. However, unclear directions from the SEA are a potential source of error in the data collected from LEAs. Terms need to be defined. For example, the difference between students served and participants is not readily clear. If the terms are not defined in the directions, it is not clear what definitions respondents will use. If different respondents use different definitions, then the data cannot be accurately aggregated.

The requirements for sustained effect studies have generated two examples of sources of error relating to the provision of directions by the SEA. In one instance, an SEA developed a new form for reporting sustained effects of first grade programs conducted in FY 1987, and then announced in a meeting with LEA representatives that no instructions would be provided. This is clearly an example of failure to provide directions. Of course, the LEA representatives inquired about details of the new report. In response, the SEA instructed the representatives to complete the report and attach an explanation of the rules the LEA followed in completing the report. This is an example of what the author has termed LEA defined direction sets. It will be difficult, if not impossible, to aggregate data from those reports.

Source of Error
° Directions for completing forms
° Unclear directions
° Absence of directions
° LEA defined direction sets
° Analysis of data
° Inappropriate statistical procedures
° Incorrect aggregation of data
° Failure to apply reasonable edits to data
° Failure to consider reasonableness of results.

Figure 2

Summary Description of Sources of Error
at the SEA Level

Once the data has been collected from the LEAs, the author believes the SEA should apply a number of edits to the data. These edits would be facilitated if the data were delivered by the LEA in computer processable form or were entered into a computer. Some edits are obvious. Certain entries should sum to other entries, column or row sums should be correct, etc. SEA personnel should know from experience other errors which could be checked by computer. Certain entries on the report have a restricted range. For example, NCE scores range from 1 to 99. An NCE score of zero, should raise an immediate flag. Experience suggests that some, but not all, of these edits are made.

Once the data has been edited, errors can occur in the aggregation of data across LEAs to get SEA level data. For example, in calculating the average NCE gain across LEAs, a weighted mean, using the number of students included in the mean for each LEA, is in order. Calculation of an unweighted mean across LEAs will most likely skew the results in one direction or the other.

Once an analysis is completed at the SEA level, the reasonableness of the resulting data should be assessed. For example, one SEA level report which the author has seen, shows the average gain in reading in NCEs ranged from a -1.0 NCES to 38.1 NCES with a median of 14.5 NCES across a group of LEAs. The gain of 38.1 NCES seems unreasonably large and should be a flag to recheck one's calculations or to go back and review the student level data reported by the LEA.

The application of inappropriate statistical procedures to a set of data is not a new phenomena. It suffices to say that with the availability of statistical packages like SAS and SPSS the potential

for this type of error is increased. Unfortunately, computer routines do not, and probably cannot, check the appropriateness of analysis procedures applied to a given set of data.

Summary

The sources of errors described above are a combination of actual errors observed by the author and scenarios contrived by the author. Four sources of error originating in the LEA were described - improper scoring of tests, improper reporting of scores, lack of attention to test fit, and lack of formal evaluation training. Seven sources of error originating in the SEA were described - unclear directions, absence of directions, LEA defined direction sets, inappropriate statistical procedures, incorrect aggregation of data, and failure to consider reasonableness of results.

The author is of the opinion that two actions can help to reduce the effects of these sources of error on the data. First, an ongoing effort to increase the level of skills in evaluation and testing of many of the persons involved in the reporting system at both the LEA and SEA level should help reduce the errors from these sources. Second, the use of computer technology, readily available in education today, permits editing of data by the LEA and SEA which should reduce errors.

SELECTING CHAPTER 1 (COMPENSATORY EDUCATION) STUDENTS USING MULTIPLE SELECTION CRITERIA

Sandra Pakes
Psychological Corporation
(Formerly with Advance Technology)

(Note: The Chapter 1 law has been reauthorized with the provisions requiring for negotiated rulemaking for setting the standards for certain sections of the law. The new law requires the establishment of National Evaluation Standards, based on the comments heard and compiled at each of the five regional meetings this summer. A selected group of Chapter 1 individuals and interest groups will meet later this summer to come to consensus on the issues and suggestions. Keeping this in mind, this paper discusses the alternatives to student selection, using multiple selection criteria.) The Chapter 1 law has previously required that a uniform method of student selection be used in determining participants for Chapter 1 Compensatory Education programs. This did not necessarily imply that a standardized test be used. It did mean, however, that a standard set of procedures be used within each grade level and between campuses that provide Chapter 1 services. The regulations stipulated that students with the greatest educational needs be served by the compensatory education programs. This meant that student selection became a major step in the process of implementing the federally funded compensatory education programs.

Chapter 1 students election must be aligned with the needs assessment information. An objectively-based, systematic process must be used to select the students in greatest need of remediation. The selection plan must be carefully planned by a representative group of Chapter 1 professionals: Chapter 1 staff, administrators, and classroom teachers. The plan should remain practical and the data sources should be objective measures. The student selection process must generate a prioritized list for each subject area served in the program and students must be selected in order from the prioritized lists.

There are a number of basic guidelines for Chapter 1 students selection which need to be kept in mind while making decisions for student placement. Students greatest in need are selected in order according to the prioritized list.

- Teacher referral alone is never adequate for placing a student into the Chapter 1 program.
- Objective information, e.g., test scores, basal placement, grades, must constitute the majority of the selection process.

- Non-placement of high priority student or placement of a low-priority student instead of a high high priority student must be documented.

- Acceptable reasons for non-placement of high priority students may include: parents refusal for placement, student's needs addressed through other services, scheduling difficulties.

Students can be selected in a variety of ways, making use of one type of information or more than one source. Often, however, using a single piece of information, such as an achievement test score, does not provide a complete picture of the student's needs. It is for this reason that composite scores, a combination of various types of information, are often used. In general, student selection should be

- OBJECTIVE, in order to ensure that decisions about each student are made on the same basis, using the same criteria for each student, and

- SYSTEMATIC, in order to ensure that all of the pieces of information are put together in a planned, logical manner.

There are many possible elements to use in a composite score as there are ways to describe a student's ability and performance. These elements, or types of information, can generally be divided into three categories: achievement test scores, teacher judgements of student performance, and other student performance indicators. Examples of achievement test scores include norm-referenced tests, criterion-referenced tests, diagnostic tests, competency tests, and end of unit tests supplied by publishers. Teacher judgements of student performance should be based on objective information such as specific skill deficiencies, relative position in class in regard to subject areas, self-concept measures, and overall need for Chapter 1 services. Other student performance indicators may include information from other data sources such as report card grades, attendance, grades behind in school/grade retention, and daily or weekly class work.

There are a number of options for establishing the district prioritized lists. Select an option which will be the best process for determining which students in the district should be placed in the Chapter 1 program, remembering that the selection criteria can vary across grades within a district but must be consistent within grades. The planning committee must also establish an appropriate cut-off score for the ranking. There should be a minimum number of points a student must compile in order to be selected for the program. The cut-off score must be high enough to require more than teacher referral

and limit the number of Chapter 1 students to the case load approved in the Chapter 1 application. In order to facilitate the process, the district should establish equivalent criteria across grade levels, keeping the same possible total points and number of criteria used. Teacher referral, if included in the selection process, must be a written referral.

TEST SCORE ONLY. The student with the lowest test scores is the first selected into the Chapter 1 program. The student with the second lowest test score is selected next. The advantages to this method are obvious. The selection process is simple and objective-based. However, student's test performance is not often the best indicator of classroom performance. A single score may not reflect overall student achievement.

TEST SCORE COMBINED WITH TEACHER'S REFERRAL. Using this option, the test score and the teacher's referral are combined in a systematic manner according to the indicated severity of the student need. Point values are assigned to the teacher referral. Selection is based on a total or combined score.

WEIGHTED LIST. A combination of multiple criteria is generally felt best in order to provide a balanced reflection of the student's needs while maintaining reasonable time and effort requirements from teachers and administrators. Three or more criteria may be combined. These may include a test score, teacher's referral and an indicator of student performance. The selection may be based on the total or combined score.

COMPOSITE SCALE. This method combines four or more criteria for a total score. It is commonly recommended that a maximum of five criteria be included in a composite scale for reasons of time and practicality. Criteria may include test score, report card grades, teacher referral, retention in previous grade(s), previous placement in a special program, parent/student request, absenteeism, or diagnostic test scores. Selection is based on the total score. All criteria included on the composite scale must be gathered for every eligible student.

When using the composite score scale, the following steps may be helpful in determining the composite score.

1. DECIDE ON THE ELEMENTS TO BE USED IN MAKING UP THE COMPOSITE SCORE. This will depend, in part, on what information is already generally available about the students or on what information can be obtained in time. This decision will also be affected by the preferences of the individuals using the composite score and the credibility they give to the various categories of student information.

2. DEVISE A NUMERICAL SYSTEM FOR SCORING EACH ELEMENT

CHOSEN. A word of caution is necessary here. Be sure that the numerical systems are compatible. For example, if you are using a composite of an achievement test and a teacher judgement scale, be sure that both numerical scales go in the same direction. The lower the achievement score, the greater the need for the Chapter 1 program. The teacher judgment scale, in order to be compatible, must use the lowest number to indicate the greatest need for Chapter 1 and the highest number to represent the least need.

3. FOR EACH STUDENT, OBTAIN THE MEASURES OR SCORES FOR EACH ELEMENT IN THE COMPOSITE SCORE. Depending upon the types of information being used in the composite score, from existing school records. Examples of these kinds of measures would include past grade, attendance record, or scores on tests previously administered. On the other hand, the measures may have to be collected. For example, the Chapter 1 teacher may want to use a rating scale to assess all potentially eligible students. Or, perhaps a certain test will be administered. In these cases, new data are being obtained for including in the composite score.

4. WEIGHT THE SCORES, IF DESIRED. Remember that the purpose of weighting a score is to cause it to have more influence in the overall student ranking. Be sure to consider how much you want that particular score to affect the ranking when using weighted scores.

5. ADD EACH STUDENT'S SCORE TOGETHER TO DETERMINE THE COMPOSITE SCORE. Be sure to be accurate when calculating the total composite scores. A careless error can seriously affect a student's ranking. It is a good idea to have a second person independently calculate the totals, so that accuracy is insured.

6. RANK THE STUDENTS IN TERMS OF "NEED FOR CHAPTER 1" BASED ON THE COMPOSITE SCORES. Be sure that the student with the greatest need is placed in the first position, and then work down to the last ranking which indicates the least need for the Chapter 1 program. Again, it is a good idea to have another individual rank the scores independently and compare results for accuracy.

7. ASSIGN STUDENTS TO THE CHAPTER 1 PROGRAM, BY RANK, UNTIL ALL AVAILABLE PLACES ARE TAKEN. Once the students are listed in rank order, it is easy to count down from the student listed in the first position until all available spaces in the program are filled.

So far, the discussion has provided examples and options in which the same type of information has been available for all of the students being ranked. But, that is not always the

case. Part of the data utilized in the composite score may be missing for a variety of reasons, such as the student changing schools, or absent on a test date, etc. There are two suggestions of ways to compensate for any missing information.

The midpoint of the missing data can be substituted. For example, if the missing data ranges from 1 to 10, the number 5 would be used. This suggestion, however, is not always the best solution. If most of the students' scores are very low, then using the midpoint as a substitute for missing data would give the student an artificially high score. Similarly, if most of the students' numbers were very high, as might be the case if days absent were counted, then using the midpoint could again be inaccurate.

If you feel the midpoint would be inaccurate, calculate the average number for those students on which you do have the information. Then, using this method, the actual average is the missing data. For example, assume that each student's grade in math for the previous year was being used as part of the composite score, but, this information is missing for some of the students. The teacher could decide to use an average grade of C for this missing data. The teacher may not think this is a fair estimate of the missing data for the student, since most other students for which there is data have math grades of D and F for the previous year. So, the teachers decide to calculate the average grade, based on the existing data (the actual math grade for last year). This calculated average turns out to be a grade of D, not the C grade.

As the above discussion indicates, student selection criteria can be very flexible and situation specific. The overall goal is to objective and systematically make use of a number of types of information through the calculations of composite score in order to determine which student should be included in the Chapter 1 program. The selection plan should be uniform within grade levels across schools. Multiple selection criteria take several factors related to school achievement into account into making these placement decisions.

SETTING CHAPTER 1 STANDARDS: A CONTINUING DILEMMA

Robert J. Nearine

Hartford, Connecticut Public Schools

Overview

How does a school district measure program success? And what should be its standard? While these questions have been raised during years of Chapter 1 operations, changes in the enabling legislation continue to leave answers to these questions generally unresolved. Current legislation (PL 100-297) continues the requirement that local educational agencies (LEA) regularly evaluate and report Chapter 1 program findings so that these can be aggregated and reported as part of a national picture of Chapter 1 accomplishments, but mentions neither appropriate evaluation models nor applicable standards. Despite no specific legislative guidance this paper suggests that the setting of realistic local standards which consider individual project service patterns and goals is a key element in the Chapter 1 evaluation process.

Background

During its twenty-three year operational history (1965-88), both critics and supporters have raised questions about Chapter 1 program effectiveness. In all probability, these were occasioned in part by a lack of specificity in the initial legislation (Title 1, ESEA as amended). While Title 1 program evaluations were mandated, the requirements were generally vague; the evaluation profession itself was still in its infancy. As educators learned more about evaluation, and become more sophisticated in their approaches to questions of accomplishment, evaluation practices improved despite the fact that regulatory requirements were not particularly rigorous. Even today, the Chapter 1 law only requires that Chapter 1 programs be evaluated with respect to their effectiveness in achieving program goals, include objective measurements of educational achievement in the basic skills, and consider program evaluation results in planning for program improvement (PL 97-35, as amended).

The concern for high quality comparable Chapter 1 data is not a new one. The U.S. Department of Education (ED) commissioned several studies which were designed to meet national as well as local decision-making needs. To obtain compatable test data, the Anchor Test Study (Loret, et al., 1974) equated nationally used standardized achievement tests, while the Title 1 Evaluation and Reporting System (TIERS, Tallmadge & Wood 1976) was developed to provide "...meaningful, comparable information about Title 1 projects at the ...building..., district, State and Federal levels (p.1)". Most recently, the Gap Reduction design was developed, again under an ED contract, for use with bilingual education and other compensatory education programs (Tallmadge, 1988) to provide meaningful and comparable information to decision makers.

There was a common theme embedded in both the TIERS and Gap Reduction models. The TIERS model assumes that with no compensatory, or supplementary assistance, the percentile rank of a group will not change over time (Model A). The model also presumes that these compensatory services can make a difference, and if these services are effective, the percentile rank of the treatment group will improve. The Gap Reduction design states and the TIERS model infers that effective programs can reduce the gap between project students and their non-project peers; at least these programs will help to keep project students from falling further behind non-project counterparts.

TIERS contained several alternate models for evaluating Chapter 1 programs. While the choice was left largely to local districts, the use of TIERS was mandated by regulation. With the 1981 enactment of the Education Consolidation and Improvement Act (ECIA), and its subsequent amendments in 1983, Title 1 become Chapter 1, but with fewer regulatory requirements. TIERS was no longer required.

Chapter 1 retained the Title 1 requirement that districts evaluate their programs at least once every three years and assess whether performance gains were sustained. While Title 1 required districts to use the federally-developed TIERS evaluation model, ECIA prohibited any federal regulations which related to "... the details of ... evaluating programs and projects (Section 591 (b)). While this legislative change allowed the states to take any approach to evaluation which they deemed desirable, a majority of both states and districts continued to use the evaluation models which had been developed under Title 1 (Birman, et al., 1987). Although additional student demographic data along with annual test data were collected, the TIERS evaluation procedures were generally continued, and in spite of reduced federal requirements. Even so, problems in the technical quality and completeness of recent evaluation data have been reported, while the extent to which districts use evaluation results to improve their Chapter 1 projects may have declined (Reisner & Marks, 1987). By themselves, current federal evaluation requirements do little more than assure that some evaluation is done and that the resulting information is available for decision making, if the local staff chooses to use it (Birman, et al., 1987).

Future Directions

While the mandated use of TIERS set a precedent for at least considering a common approach for the collection and reporting of Chapter 1 evaluative data, the same precedent also argues persuasively for the establishment of local district and project-level standards for success. This precedent can be explained in response to the following questions.

1. Should one overall standard be established? Hartford, and perhaps other districts, recognize that a set standard for all of our Chapter 1 programs is probably not reasonable. Each program differs in the nature of its clientele, and expecially in the land, duration, and nature of the services which are provided. For example, the same project impact cannot be expected from a short summer school project as would be expected from a full year program. For secondary school youngsters who have shown patterns of failure through elementary school, and have in all probability been falling continually behind the norm group, gains of 0 NCEs may be optimal; at least the youngsters are holding their own. And, for an elementary remedial program, a 5 NCE gain on

average could be expected but only if the youngsters were in regular attendance.

2. How much gain should be expected? In the original TIERS Users Guide (Tallmadg & Wood, 1976), the authors discuss the use of NCEs to assess gain and the establishment of a standard setting procedure. While no specific number of NCEs is mentioned, the authors point out as a rule of thumb, that when treatment and control groups are found to differ by more than 4 NCEs (or a fifth of a standard deviation with respect to the national norm) other models may be preferable (page 25). This is the only place when a number is mentioned. The establishment of a target gain level is left to the local district and/or to a given project. Further review of compensatory education handbooks, both current and present, and materials which were provided both by one SEA (Connecticut) and the regional TAC over recent years (e.g. RMC materials, etc.) provide essentially the same information; while there were no established goals, districts/projects were encouraged to set standards which were statistically and practically defensible. Since the Gap Reduction design also emphasizes the use of local comparison groups so as to make results more easily interpretable, both TIERS and the Gap Reduction model make the point that standards for success essentially should be local project level decisions.

At the same time, TIERS user-districts were also concerned about how much growth is good and consequently, Connecticut also provided an unofficial target level of 5 NCEs; about a fourth of a standard deviation (SD). This figure and various research articles suggest that while both 7 NCEs (one-third SD) and 4-5 NCEs (one-fourth SD) are reasonable levels of expectation, changes of only 1 to 3 NCEs on a one year basis may only reflect ideosyncratic and unimportant changes in the measurement. Small changes such as these should best be viewed over time.

Conversely, at least one researcher has argued that gain standards of 1 to 3 NCEs should be established of a state-wide basis (Fitzgerald, August 1986). While these minimal standards may appear at first to be conservative levels of expectation, even these low overall standards suggest potential problems. Gain standards of 1-3 NCEs may not be defensible to peers, superintendents, or Boards of Education unless there is a clear basis for using this apparently low standard. Since many Boards of Education and educators have been sold on the concept that normal gains can be attained in the classroom while above normal gains require supplementary services, higher expectations are effectively negated by using one limited standard. Note also that since a 4-7 NCE gain corresponds to several months grade quivalent gain (above and beyond year for year growth) and also equates to about one half of a stanine, changes of this magnitude can be seen as being important, while lower standards seem to reflect minimal growth.

3. What direction should the Chapter 1 regulations take? While it is anticipated that the regulations will remain somewhat flexible, there are several concepts which these should contain.

a. Districts/projects should be encouraged and helped to set high but realistic levels of expectation for their projects. This encouragement is apparent, if not explicated, in several of the older TIERS publications, and is emphasized in the Gap Reduction design.

b. Technical Assistance Centers (TAC) and states could assist districts to develop reasonable levels of expectation by analyzing NCE gain data obtained from students who are not currently receiving compensatory services. In states where mandated testing programs are in operation (e.g., Connecticut Mastery Test), a state-level metric could be used as a common data source. While the TAC/state might specify program categories and suggested gain levels based upon various service patterns, no one level should be set for a state overall.

c. The adequacy of gain patterns should only be judged in the context of a given program. Should there appear to be problems with a project, state and TAC personnel can/should be encouraged to help districts initiate corrective actions but only on a bilaterial basis and after an assessment of all available facts. The state should not issue "go/no go" decisions on the basis of a lack of NCE gain scores alone.

Quality evaluations help everyone; the learners and the decision makers as well. A large measure of this quality is dependant on the establishment of appropriate standards; these should be rigorous, attend to target group needs and local project service patterns, and should be used as one criterion of program success. The upcoming Chapter 1 regulations should address these two concepts by emphasizing rigor which responds to local needs.

REFERENCES

Birman, B.F., Orland, M.E., Jung, R.K., et al. (1987). The current operation of the Chapter 1 program. Washington: Office of Educational Research & Improvement, U.S. Department of Education.

Connecticut State Board of Education. (1986). Compensatory education handbook. Hartford: State of Connecticut Department of Education.

Fitzgerald, N.B. (August 1986). Program improvements through evaluation: Connecticuts first round of sustained effects studies. Feedback, Occasional paper series (Paper #1). Hampton, NH: RMC Research Corporation.

House of Representatives. (April 13, 1988). Elementary and secondary education conference report to accompany H.R.5. Washington: U.S. Government Printing Office.

Tallmadge, G.K. (1988). Implementing the gap reduction design. Handout, AERA Mini-Training Course 3.37, April 5, 1988.

Tallmadge, G.K., & Wood, C.P. (1976). Users guide ESEA Title I evaluation and reporting system. Mountain View, CA: RMC Research Corporation.

The Evaluation of Migrant Programs

Patty Higgins

Georgia Department of Education

The mobility of students, reduced number of days in attendance, and other characteristics make the goals of the migrant program focus on things other than reading and math achievement. Supportive services are designed to reduce the negative effects of being migrant on the students. Their goal may be to increase attendance to enable the students to benefit from the regular program. They may be tutorial to assist the student in keeping up with their regular program (not measured through reading or math subtests). They may be to reduce the dropout rate by increasing success and confidence within the regular program. All of these are not measurable on tests but may be measured in other ways.

The individualized nature of the migrant program and the mobility of students does not lend itself to standardized norm referenced evaluation. The best procedure for national evaluation is through the establishment of national objectives for the program that can be measured. Appropriate data could be collected from all programs. Examples of these are:

Number of classroom attendance days
Number of students passing at the end of the year
Number of students graduating from high school
Number of students taking the SAT and their scores
Student grades for each subject taken
Length of time it takes for student records to be received, etc.

Because several states may serve individual students within a year, any pre/post test using TIERS cannot be used to evaluate a SEA migrant program. However, if deemed necessary, it may be used to evaluate the impact of the migrant program nationally. The collection of such data could be done through the MSRTS system with all states entering the percentile or NCE score of students. A computer program could be written to aggregate data from the MSRTS data base. A second alternative would be to use the NAEP program results with an oversampling of migrant students. This would meet the requirements of the law for the program to be evaluated but would not impose excessive burden on LEA or SEA personnel. Since the reliability of the data is questionable and probably does not actually evaluate the impact of the program, it would simply be a paper process to satisfy the law. If this is the case, it should be done with as little burden as possible.

Using Equated Norm-Referenced Test Scores for Chapter 1 Reporting Purposes

Sharon Johnson-Lewis
Detroit Public Schools
Detroit, Michigan

In the Spring of 1984 the Detroit Public Schools initiated an innovative citywide testing program. The program consisted of the administration of both norm- and criterion- referenced tests. The norm-referenced, California Achievement Tests (CAT), Form C, were to be administered to all students in Grades 3, 5, 8, and 11 in the fall of each year and the locally developed criterion-referenced tests, Assessment of Basic Curriculum Skills (ABCS), were to be administered in the spring to all students in Grades 1-8.

The CAT was used to generate test score data at strategic intervals so that Detroit Public Schools students' achievement levels could be compared to that of a national normative group. The ABCS was used to gather information about Detroit Public Schools students' mastery on essential learning skills which had been identified by Detroit teachers and curriculum staff.

A primary concern of Detroit staff was the dual testing of students for the citywide testing program and for Chapter 1 reporting purposes. In order to alleviate that problem, Detroit staff worked with Dr. Benjamin Wright, University of Chicago, and Dr. Susan Bell-Masterson, RASCH analysis specialist, to equate ABCS test results to the CAT.

Detroit staff obtained permission from the test publisher, CTB McGraw-Hill, to include CAT items with the field testing of the ABCS test items. CAT test items were used as "anchor" items for each of the over 100 field test booklets. Best Test Design, co-authored by Benjamin Wright, was used as a guide for test development. Based on data provided by Dr. Wright and Dr. Bell-Masterson, tables were constructed equating the California reading and mathematics subtests, respectively, to the Assessment of Basic Curriculum Skills. In order to validate the equating process students were administered both tests in the Spring of 1984.

Validation of the equated scores consisted of comparing Pearson product-moment correlations against the published CAT reliabilities for each subtest administered. Pearson product-moment correlations were computed using two pairs of variables, the actual CAT scores and the CAT scores which had been derived from ABCS raw scores. In addition, it was expected that at a minimum, a Pearson product-moment correlation would not be more than 10 points (0.10) smaller than the publisher's reliability for the corresponding CAT

subtests. This goal was obtained for all five of the reading comprehension subtests and in 2 out of 5 grades for the mathematics total score.

The derived CAT scores were used for Chapter 1 reporting purposes only. The process of using derived CAT test scores for Chapter 1 purposes will continue until the 1988-89 school year. Then the district will eliminate the use of criterion-referenced tests. Beginning in the Spring of 1989 the California Achievement Tests, Form E will be administered to all Detroit students in Kindergarten through Grade 12. Thereafter, the need to use equated CAT scores will no longer exist.

In the five years that Detroit has utilized equated norms many advantages and disadvantages were apparent.

Advantages

1. Utilization of equated norms eliminated the need for dual testing. Hence, a district can administer a criterion-referenced test as part of its citywide testing program and then use the equated norm-referenced scores for Chapter 1 reporting purposes.

2. School districts can administer criterion-referenced tests that are tailored to their curriculum and not depend primarily on publishers' norm-referenced tests.

Disadvantages

1. Equated norm-referenced test scores were not provided Students, staff, and community constantly sought these data. These groups need to be educated about the values of criterion-referenced scores and the limitations of norm-referenced scores.

2. In order for the equating to remain current, the procedure must be repeated on a routine basis. This process would require dual testing at least for a sample of students.

3. School district staff must work closely with a consultant and/or become proficient in sophisticated statistical techniques. In order to benefit most and to obtain the best final product, district staff usually must accomplish both of the aforementioned.

It is difficult to counter the advantages and disadvantages of the equated norms on a one-to-one basis. In addition to stating the advantages and disadvantages of the

utilization of equated norms, there are certain conditions which have been learned and which form the basis for making recommendations.

Recommendations

School districts should . . .

1. Be sure to allow sufficient time for the equating and validating process. This will usually require a minimum of two years.

2. Allocate adequate staff to conduct the equating process. This number can range from two in small districts to four or five in larger districts.

3. Work closely with knowledgeable consultants. The methodology and technology of test equating is changing rapidly. District staff should consult with persons who devote a major proportion of their professional time working with these issues.

4. Equate the test items on a fairly routine basis. The equating link is weakened in time similar to the way national norms are weakened over time. Equating should be conducted at least every two years.

Issues and Areas for Aggregating Chapter 1 Data[a]

Gary D. Estes
Northwest Regional Educational Laboratory

Interest in obtaining aggregated data for Title I and Chapter 1 programs dates back more than ten years, prior to the enactment of the ESEA Title I amendments of 1978. At that time data were not available to answer the question of how Title I programs were doing nationally, so a system referred to as the Title I Evaluation and Reporting System (TIERS) was developed and mandated. It required states to collect data from local districts, aggregate those data for the entire state, and report the aggregated results to the Education Department (ED).

The initial system consisted of three evaluation models. Within each model there was an option to use either nationally normed tests or tests without national norms. The intent was that data could be aggregated across the different models, tests, and testing cycles. But, the different models and testing cycles produced sufficiently different results that it was not desirable or reasonable to aggregrate across them. Also, aggregating across models ultimately was not an issue since nearly 100 percent of the states and districts used only one model, i.e., Model A, the norm referenced model. Differences in the testing cycle were handled by not aggregating fall-spring data with annual spring-spring or fall-fall data. These were reported separately.

Finally, the initial system included the reporting and aggregating of outcome data in the form of achievement test information, program participation data, and data on specific program characteristics such as student:instructor ratios; number of hours of special instruction; and program setting in terms of laboratory, pullout, etc. The reporting of these program characteristics data was discontinued due to a combination of factors. First, districts and states interpreted reporting guidelines differently. Second, there were significant differences in the programs that made uniform reporting difficult. For example, some states and districts would average across buildings or program components in reporting these data, whereas others would keep them separate. Third, ED, SEAs, and local districts' staff and resource limitations resulted in little analysis and use of the data.

The most recent Chapter 1 aggregation practices have focused primarily on reading and math norm referenced scores in grades 2-12 to estimate student impact. Data are maintained separately by grades, subjects, and testing

---

[a] Gary Estes is Director of Assessment and Evaluation and the Region 4 Chapter 1 Technical Assistance Center at the Northwest Regional Educational Laboratory. The views in this paper do not necessarily represent those of the Northwest Regional Educational Laboratory or the United States Education Department that sponsors the Chapter 1 TACs.

cycles. Program participation data are aggregated across districts and states, and yield information on the numbers of students served by grade, subject/service area, ethnic category, and gender.

Under the new legislation for Chapter 1, many changes have been made that raise new questions about the validity and usefulness of current evaluation practices. This paper addresses issues related to aggregating data in conjunction with evaluation and reporting options for Chapter 1 programs under H.R.5.. Below is a brief outline of the purposes aggregation can serve. It is followed by a discussion of (a) types of Chapter 1 programs for which aggreration of data might be an issue, (b) the level at which data could or should be aggregated, and (c) how different outcomes and other program variables could be included or excluded in any aggregation.

Purposes for Aggregation

Aggregation refers to the process whereby data across subelements are collapsed into a total. These elements include:

a. reporting units such as schools, districts, and states

b. student subpopulations such as by gender, ethnicity

c. different methods for producing data such as different evaluation models, measures, or cycles, e.g., fall-spring and annual data collection cycles

d. different program types such as teacher vs. aid programs or pull out vs. inclass programs.

The degree to which aggregating across units within elements is desirable depends on the purpose the information is to serve and what level of use is to be made. In general, the purpose of aggregating across units is to obtain a more global or total picture than if data are reported in the subunit form. Often, a "total" picture is appropriate and of most interest when the focus is on how the total program or system is functioning or when questions are more in a yes/no form. If the data are to serve purposes such as determining whether the program works equally well for all types or categories of students, if one type of program is relatively stronger or weaker than another, or if separate units such as particular schools or districts are meeting some standard, then having data that are not aggregated across these units is desirable.

Aggregation in Different Programs

Regular Program. The current aggregation for student outcomes is done only for the reading and math components of Chapter 1 programs. Nearly all states and districts with Chapter 1 programs are experienced in aggregating and reporting these data under the current system. Although language

arts/writing data are available in some districts and states, they are not part of the national aggregation. This is probably primarily because of the relatively small number of students served in these areas compared to the number in reading and math.

There is no universal agreement that data aggregated across districts and states are sufficiently valid and reliabile; but the results of Chapter 1 and Title I data from these aggregations are fairly consistent with the results from other more controlled studies such as the Sustaining Effects Study conducted by Systems Development Corporation or the results from the National Assessment of Educational Progress. Thus, some evidence exists that results from local and state aggregated data are valid.

Migrant and N or D Programs. Each of these programs presents some common problems and some slightly different ones. Historically, they have not had achievement data aggregated in the same manner as the Chapter 1 regular programs. They have been able to routinely report data on variables such as number of students served and, on occasion, on the achievement status of participants at a point in time. High student mobility is a common problem for aggregating data in migrant and N or D programs. Since students in these programs often are not in a school or district for as long as seven months, it is not possible to collect and match data for a school year. There are, however, a substantial proportion of migrant students that remain in a school or district for a school year or longer. Thus, the Migrant programs could aggregate data on students that remain in a school or district over a period of time if data on this subset would be of interest and use to district, state, or federal staff. On the other hand, N or D programs have relatively fewer students that remain in a school for as long as 7-9 months, so it will be relatively difficult to evaluate student progress in these using data matched over time.

Early Childhood Programs. Programs at the first grade level and below also have characteristics that make aggregating data more problematic than for the reading and math programs at higher grades. The problems can be attributed to characteristics and interactions between the students and measures at this level. First, younger children's achievement is much less reliable than older children's. Second, and in part a function of the rapid changes in children's skills, the tests at these early levels do not have sufficient reliability and differ in their structures so much that aggregrating data and evaluating in the pre-post mode is much more problematic.

In summary, currently it does not appear that Chapter 1 migrant, N or D and early childhood programs are as readily able to aggregate data as are the reading and math programs in grades 2-12 for which data have been aggregrated and used for nearly ten years. Collecting individual district or school program evaluations for migrant, N or D, and early childhood programs and conducting secondary analyses of these could provide a basis for describing the range and types of outcomes without attempting to aggregate the data.

Level of Aggregation

Data can be aggregated across students, schools, districts, and states. In fact, the current Chapter 1 system aggregates across all of these units. Clearly, when data are aggregated over units with significant variations, the "average" alone does not provide a complete picture. It is almost too obvious to state, but the degree to which data aggregated across these units is useful depends on the purposes the data are to serve. If the question is how does the program work in the state, district or nation, then data aggregated to these levels accompanied with information about variability within these units will be appropriate. When the questions are how well are the separate entities within the units performing, then maintaining or having disaggregated results is necessary.

Currently, some states collect data at the student or school level. More states, however, collect and aggregate data at the district level. Few states have used student or school level data to determine the types of assistance or sanctions to offer schools. School level data will become increasingly important with the newly passed Chapter 1 legislation in which a strong emphasis is given to school level accountability and performance.

Another form of aggregation is by categories of students. Increasingly, schools and districts are attempting to insure that all student subgroups, e.g., different ethnic groups, make satisfactory progress. Thus, dissagregated data by subgroups of students will probably become more rather than less important at the school and district levels. It is not clear that there will be a high need or priority at state and federal levels for data dissaggregated by these student subgroups. The state and federal roles might be to support and encourage districts or schools to examine their data in disaggregrated forms.

In summary, data are needed for the level at which decisions are to be made. While the need for aggregated data remains, it appears that there is an increasing need for disaggregated data at the school level to insure that individual schools and particular students all make sufficiently positive progress. Whether these purposes are met by aggregating and reporting data at higher levels or by having lower levels assure that they are examining data to address these questions, will need to be resolved in the form of policies and supporting directions from the federal and state levels.

Types of Data

Much of the discussion above focuses on aggregating student achievement data. When evaluating Chapter 1 programs, however, other types of data may also be relevant. Three categories can be considered when contemplating what types of data should or should not be aggregated: (a) higher vs. lower level-basic skills achievement data, (b) outcome data other than achievement test results such as percent of students graduating from a program; and (c) non-outcome data such as program/project and student characteristics.

In the new Chapter 1 legislation, it is clear that the intent of Chapter 1 programs is to assist students in achieving the basic and advanced skills contained in the programs offered all students. How achievement of advanced and basic skills is defined and measured is open to question. At least two options exist. Most districts or states use norm referenced test batteries or similar objective tests to assess regular student achievement. Using these same measures to evaluate Chapter 1 students progress could be viewed as assessing both the basic and advanced skills of the regular program. Another view is that only subparts of these measures tap the more advanced skills such as reasoning, critical thinking, etc., and that these subparts should be aggregated and reported separately. A third option is to mandate the use and reporting/aggregating of separate tests such as higher order thinking skills tests or performance type tests to evaluate advanced skills. Within these options it is possible that results could be aggregated across units such as schools, states, etc., without aggregating across basic and advanced skills. That is, districts could report an overall result and a result on the advanced skills subset of the overall result. Aggregating these data across schools, districts, etc., would imply a purpose of either assuring accountability or determining types of assistance and changes that are needed.

Other Outcome Measures. Data such as the percent of students who achieve well enough to "graduate" from the program could be used as outcome indicators to supplement achievement test data. As these alternative measures are developed or proposed, key questions to keep in mind include: (a) what will be the definitions and procedures needed to implement these measures; (b) what effect will variations across units such as schools, districts, and even states have on the implementation and measurement of these outcomes; and (c) where will the data be most useful and needed, i.e., what will states need vs. districts vs. schools. Again, the answers to these questions will depend on the policies that establish the role among these different levels.

Project Characteristics. Measures of project characteristics include factors or data such as the numbers and types of students served, types of programs in which these students participate, and characteristics of the program. The prior Title I reporting included student:instructor ratios, project lengths, hours of instruction, etc. We also know that variables such as academic learning time and amount of direct instruction are related to student achievement. What is less clear is the degree to which these types of data can be efficiently and accurately aggregated across levels. The degree of variation in school, district, and state interpretations and reporting of similar variables in the Title I system suggests that even with standard definitions, the variations in the ways programs are designed and implemented will make it difficult to aggregate these types of data across projects, schools, districts, and states.

In summary, expanding aggregate data to include multiple measures of student outcomes, higher and lower level skills, and descriptions or measures of program characteristics will raise key questions, such as who will use the data and for what purpose. The answers to these questions will be determined largely by the role and relation among federal, state, and local agencies that are reflected in policies. Some of these are implied in the newly passed legislation, others appear to be open to determination by federal, state and local staff.

Summary

The question of what data could or should be aggregated for Chapter 1 programs has several responses depending on the the subquestion. Some of the issues and possible responses are summarized in the table below.

Issues in Aggregating Chapter 1 Data

Issue Area	Sample Options	Option Implications
Type of Program		
Regular	Aggregate Reading and Math	Data continuity maintained, burden minimized
Migrant and N or D	Aggregate participant data but only collect samples of local districts evaluation reports	Nationally representative data may not be available on student outcomes
Early Childhood or Grades	Same as Migrant	
Level of Aggregation		
Student	Maintain at school/ district levels	Assumes that schools and districts are primary users of student level data
School	Report school level to state	Enables states to monitor individual school progress and to target assistance
District	Aggregate and report to state	Provides district level information for state monitoring and reporting to federal level
State	Report state Aggragate to ED	Provides for a national picture of the program and can be used to account to Congress or others

Issues in Aggregating Chapter 1 Data--Cont'd

Issue Area	Sample Options	Option Implications
Type of Data		
Higher level skills	Use regular tests/ measures and total reading, math, etc. scores	Assumes that these measure both the basic and higher level skills covered in the regular curriculum
Other outcomes	Require only at district and school levels	Assures local staff use information on a broad set of outcomes and not just on reading and math test scores; allows for variability in measures, definitions used by each district/school
Project/student Chacteristics	Aggregate number of students at state and federal level but do not disaggregate achievement by student categories at the state and federal levels	Provides descriptive information, but does not allow for subgroup comparisons at state or national levels

The sample options are not intended to be exhaustive. They are intended to illustrate that there are implications for any option chosen. Often the implications are along continu like (a) degree of burden, (b) level of use, and (c) range of information. In examining options for the Chapter 1 programs it appears that a maintenance system would continue the current Chapter 1 evaluation and reporting system. To respond to new mandates for school level results, states and federal agencies could simply require that districts monitor the performance in their schools. A more expansive, revised model could include (a) reporting school level data to state agencies; (b) implementing standard procedures for reporting achievement data for migrant, N or D and early childhood programs; (c) requiring separate aggregations and reporting of basic and advanced skill data; and (d) disaggregating results by ethnicity, program type, etc., and reporting these results from the school to district to state to federal levels.

As TIERS was implemented in the late seventies and early eighties, USED staff and contractors, school district evaluators, and researchers examined how the system worked when implemented over all states with different programs, staff, etc. This produced useful information about the strengths, weaknesses, and feasibility of the data and reporting systems. Hopefully, any Chapter 1 evaluation expansion to new areas, application of new procedures, or collection of new data will be accompanied by similar research and examinations before final decisions are made about their feasibility and utility.

In closing, it is important to note that mandating evaluations and reporting of results does not automatically insure use of these data to improve programs. A commitment by and resources to assist local, state, and other agencies to use whatever data are available will facilitate the use of these data for improvement.

Comparing TIERS Model A with the Gap-Reduction Design

G. K. Tallmadge
RMC Research Corporation

TIERS Model A and the Gap-Reduction Design are closely related. In fact, TIERS Model A is a special case of the more general and more widely applicable Gap-Reduction Design.

In the Gap-Reduction Design, the amount of gap reduction is defined as the standardized pretest gap (between the project group and the comparison group) minus the standardized posttest gap (between the project group and the comparison group). Algebraically, this is:

$$\text{Gap Reduction} = \frac{\bar{X}_c - \bar{X}_p}{SD_{xc}} - \frac{\bar{Y}_c - \bar{Y}_p}{SD_{yc}}$$

where:
- $\bar{X}_c$ = the mean pretest score of the comparison group
- $\bar{X}_p$ = the mean pretest score of the project group
- $\bar{Y}_c$ = the mean posttest score of the comparison group
- $\bar{Y}_p$ = the mean posttest score of the project group
- SD_{xc} = the standard deviation of the comparison group's pretest scores
- SD_{yc} = the standard deviation of the comparison group's posttest scores

If we use the 50th percentile of the national norms as our comparison group and NCEs as our test-score metric, $\bar{X}_c$ and $\bar{Y}_c$ equal NCEs of 50.

Similarly, both SD_{xc} and SD_{yc} equal 21.06 NCEs.

Substituting these values in the equation above gives us:

$$\text{Gap Reduction} = \frac{50 - \bar{X}_p}{21.06} - \frac{50 - \bar{Y}_p}{21.06}$$

This equation simplifies to:

$$\text{Gap Reduction} = \frac{\overline{Y}_p - \overline{X}_p}{21.06}$$

With TIERS Model A, project impact is defined as the project group's mean posttest NCE minus its mean pretest NCE, or

$$\text{Project Impact} = \overline{Y}_p - \overline{X}_p$$

As can be seen by comparing the last two equations, a Model A NCE gain can be converted to a gap reduction simply by dividing it by 21.06.

It also follows that, if the Gap Reduction Design is implemented using the 50th percentile of the national norms as the comparison group and NCEs as the test-score metric, multiplying the amount of gap reduction by 21.06 will yield exactly the same result that would be obtained by implementing Model A.

It is important to note here that the essential equivalence of the two designs holds only when national norms are used as the comparison group and only when NCEs are used as the test score metric. These conditions are quite restrictive. We might want to use something other than a nationally normed test to evaluate our project, or we might want to evaluate our project using non-test data such as attendance rates or classroom grades, or we might want to pre- and/or posttest at times that do not correspond to empirical norming dates. The Gap-Reduction Design can be easily implemented under any of these circumstances--but Model A is either totally unusable or usable only after performing one or more moderately complex and error-prone statistical manipulations.

The biggest advantage of the Gap-Reduction Design over TIERS Model A is that it can be used with "live" comparison groups. This feature is important because, if your project students were not in your project, they would be in mainstream classrooms in <u>your school</u>, not in some hypothetical national-average school.

When norms are used as the comparison group, you are comparing the growth of your project students' against national averages of the growth of similar students. But the growth of your project students is a function of both what they learn in your project and what they learn from non-project instruction in your school. If that non-project instruction is more effective than the national average, the extra learning that results from it will, in effect, be added to the project-

related learning of your students. Your project would then be made to look more effective than it really is.

If the non-project instruction in your school is less effective than in the nationally average school, the difference will, in effect, be subtracted from the project-related learning of your students. Your project would then be made to look less effective than it really is.

In all cases with TIERS Model A, project effects are confounded with school effects. To avoid this problem, you must use a live, local comparison group. This cannot be done with TIERS Model A but is simple and straightforward if you use the Gap-Reduction Design.

To summarize, the Gap-Reduction Design will yield essentially identical results to TIERS Model A if national norms are used as the comparison group and NCEs are used as the test-score metric. Under other circumstances, it has the following advantages:

- It can be used with non-normed tests and even with non-test data (e.g., attendance rates).
- It allows comparison with local groups, not just a nationally average group.
- It allows testing to be done whenever it is convenient, not just near empirical normative data points.