DOCUMENT RESUME

ED 344 416 EC 301 145

AUTHOR Callahan, Carolyn M.
TITLE Determining the Effectiveness of Educational Services: Assessment Issues.
PUB DATE Mar 92
NOTE 7p.; In: Challenges in Gifted Education: Developing Potential and Investing in Knowledge for the 21st Century; see EC 301 131.
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Educational Assessment; Educational Quality; Elementary Secondary Education; *Evaluation Methods; *Gifted; Outcomes of Education; Program Effectiveness; *Program Evaluation; Qualitative Research; Standardized Tests; Student Evaluation

ABSTRACT

Various issues in determining the effectiveness of educational services for gifted students are considered in this chapter. First, the limitations of standardized instruments are identified. These include narrowness of assessment (usually only across traditional curricular areas), invalidity in assessing program goals (which may not correspond to test areas), ceiling effects (an insufficient range of items at the upper end of the continuum), and regression to the mean (a statistical phenomenon which may hide actual growth in achievement). Alternative assessment strategies are then proposed such as locally developed assessment instruments, out-of-level assessment, multiple assessment, assessment in nontraditional areas or in traditional areas using nontraditional means. The issue of standards in gifted program evaluation is discussed noting the effect of national achievement standards on assessing gifted student achievement and changes in grade equivalency scores as standards. Finally, both the use of control groups to determine outcomes and the use of qualitative evaluation strategies are considered. (DB)

Determining the Effectiveness of Educational Services: Assessment Issues

Carolyn M. Callahan — *Professor of Educational Studies and Associate Director of the National Research Center on the Gifted and Talented, University of Virginia*

Introduction

Determining the effectiveness of educational services is predicated on the assumption that evaluators know what to look for as reasonable indicators of success, that they know how to assess change on those indicators and that they know how to interpret that change when it occurs. These assessment issues continue to plague evaluators, teachers, administrators, and parents involved in programs for the gifted.

Limitations of Standardized Instruments

One of the lingering problems facing those who attempt to determine the effectiveness of programs for the gifted is the selection or construction of instruments that will yield reliable and valid information. A second, related problem is the determination of meaningful indicators of success acceptable to those who must use the information for decision making.

These problems stem from several sources, one of which is the mismatch between existing standardized instruments and the goals and objectives of programs for the gifted. Existing standardized achievement instruments have been criticized for their narrowness, invalidity in assessing program goals, and potential ceiling effects.

Narrowness of Assessment

Standardized tests are designed to assess the traditional curriculum. Test developers go to great lengths to ensure that the standardized tests reflect the predominant curricular goals at the appropriate grade level nationally. This results in very specific assessment across traditional curricular areas within a narrow range of expectations. The selection of both the content and the level of thinking required to answer the questions on standardized tests reflects the aim of assessing those areas to which most students are exposed in their traditional curricula.

This stands in contrast to the broad and extended goals of programs for the gifted. These goals encompass such skills as the development of creative and critical thinking and the development of independent learning skills. In addition, the content of most gifted programs falls under the rubric of "enrichment" (content not normally included within the traditional course of instruction) or "acceleration" (content normally taught at a higher grade level). On-level standardized tests, those tests given to students based on their age, are not effective in assessing either knowledge of content or the level of thinking acquired by gifted students.

Invalidity in Assessing Program Goals

This narrowness of assessment leads to a validity issue. *Validity* is a term used to describe the degree to which a test assesses what it is intended to assess. A test may be valid for one purpose but not another. Standardized tests are an excellent example of this tenet. That is, standardized tests might be quite accurate in assessing the outcome of the traditional program because the items closely match the goals and objectives of that program, but they may be very invalid for assessing the goals of a program for the gifted. Even when the standardized tests propose to assess the thinking skills that are now being integrated into all curricular areas, caution must be exercised. Too often, the level of thinking required of the regular student is still very low and not in concordance with the objectives of gifted programs.

When gifted students are in a program based on the acceleration model, the use of standardized tests may be justified if there is a close match between the test chosen and the content to be taught. However, even when the content of the curriculum offered to gifted students is accelerated, the use of standardized assessments presents very particular problems. If on-grade assessments are used, the range of items presented to students will fail to assess the goals and objectives of the program because so few items will go beyond the traditional grade level. If out-of-level assessments are used, the new types of answer sheets, the typeface and size of type, the density of problem presentation, and even the length of the test may interfere with the assessment process unless considerable caution is exercised.

Ceiling Effects

One further problem with using standardized tests is the potential that such tests will not have a sufficient range of items at the upper end of continuum to fully assess student growth. This is a problem particularly when students are identified for placement in a program for the gifted on the basis of high scores on these tests. If a student earns a score between the 95th and 99th percentile on a standardized achievement test, that student most probably has responded correctly to nearly all of the items prior to any instruction at all. Change cannot be measured on these instruments because there are not enough items at a level advanced enough to be sensitive to that change.

Regression to the Mean

High scores on a pretest also will result in a statistical phenomenon called regression to the mean, which may act to hide actual growth in achievement. This effect occurs because all test scores contain some random error. On any given day, an individual may earn a higher or lower score, not because of what the individual knows or can do, but because that person guessed correctly, marked an answer incorrectly, or misread a question. For students who earn high scores, the error factor was in their favor that day. When they take the test again, the error may work against them this time and lower their scores (i.e., cause their scores to regress to the mean), even though they may have gained in knowledge or skills.

The Lack of Suitable Measures Addressing the Outcomes of Programs for the Gifted

Although some measures of critical thinking skills and creativity have been used in assessing the effectiveness of programs for the gifted, these measures are usually limited in scope and address only a few of the many goals of gifted programs. For example, the Torrance Tests of Creativity (TTCT) are considered reasonable measures of the specific skills of fluency, flexibility, and originality. However, many experts question whether scores on these tests are suitably generalized to an assessment of the larger construct known as creativity.

Further, many of these instruments (e.g., the TTCT, the Ross Test of Higher Cognitive Processes, and the New Jersey Test of Reasoning Skills) are used in the identification process; the students are selected for eligibility to programs on the basis of high scores on the tests. Thus, if these same instruments are used to assess change, ceiling effects and regression to the mean effects may interfere with the assessment of true change.

The number of instruments specifically designed to evaluate the goals most common in programs for the gifted is also very limited. This is not surprising in light of the complexities of gifted programs and the drawbacks of developing instruments for such programs. Test development and publication are extremely expensive and require considerable distribution

to warrant the investment. The market simply does not exist for instruments that assess the outcomes of gifted programs.

There are no "national" or "state" curricula or even objectives for programs for the gifted. Even in those areas where there might be agreement on general goals, such as the development of creativity and higher-level thinking skills, there is little commonality in definition; little agreement about the appropriateness of a given skill for a given grade level; and no common content base, except in some acceleration programs and in advanced placement courses. In many programs based on individual programming models and independent study, there may not even be a common set of goals for all gifted students served by the same program, such as a program based on individual student needs. Creating standardized assessment tools under these conditions would be a futile effort.

Alternative Assessment Strategies

Locally Developed Assessment Instruments

Locally developed assessment instruments are the logical alternative. Unfortunately, few school personnel have the expertise in instrument construction to collect appropriate evidence of reliability or validity of instruments, to establish reasonable norms and standards, and to avoid the tendency to create very content specific instruments. Finally, in programs that are individualized in nature, it is nearly impossible to identify one instrument that will assess the myriad of goals that might be part of the educational programs of even a small number of students. However, with some investment of time and effort, data from local instruments provide useful information, especially if used as indicators of success (see "Multiple Assessment").

Out-of-level Assessment

As suggested above, assessing a program that is based on acceleration may be accomplished by out-of-level testing. But out-of-level assessment should occur only when the objectives assessed by the level of the instrument selected match instructional objectives. Even when that match is clear, care must be taken at certain grade levels to ascertain that the format of the test does not interfere with performance.

For example, in testing second graders on upper-level forms of the Iowa Tests of Basic Skills, the format of answering on separate answer sheets was found to be new, confusing, and ultimately inhibiting to performance. Some standardized tests use much less "white space" on upper-level tests, and younger students may find the page layout very distracting. Also, students accustomed to being able to answer all or nearly all of the items on a test can become very frustrated with a test that has more difficult items. All of these observations suggest that careful examination of the test objectives, the format, the instructions, and the length of the test must precede use of the instrument.

Multiple Assessment

One of the tools used by evaluators to assess programs is a procedure called triangulation. This technique is used when the evaluator does not have a perfectly controlled design, a perfect instrument, or perfectly unbiased sources of information. The term "triangulation" evolves from the strategy used to locate a position or point when surveying; the essential principle is to use multiple indicators from multiple sources to assure consensus on the finding. This concept can be applied to assessing student outcomes by thinking in terms of multiple assessment strategies, multiple measures, and multiple scorers or raters.

For example, in assessing the student outcomes in a program that emphasizes creativity, the following measures might be used:

- A standardized test such as the TTCT, assuming it was not used for identification, might be used to assess some of the related skills that are specifically addressed in the program.
- A product rating scale, completed by "experts" in the discipline of the product, with a dimension specifically focused on characteristics of student products that are creative in nature may be used as one additional source.
- A process rating scale, completed by the teacher to rate the steps engaged in by the students in creative production, can provide further data on the achievement of goals.
- A self-rating scale on self-perceptions of changes in skills related to creativity might be completed by the students.

Comprehensive examination of all of these data points will generate a better assessment of the overall construct.

Assessment in Nontraditional Areas and in Traditional Areas Using Nontraditional Means

There are some disciplines in which the use of any standardized instrument, especially any standardized paper and pencil assessment, is totally unwarranted. These disciplines include the visual and performing arts and leadership. Although a growing number of individuals would argue that all paper and pencil assessment is artificial and uninformative, regardless of the discipline, the development of portfolio and performance assessment tools is critical to gain an authentic assessment in the arts and leadership areas.

Assessment tools can relate to *processes* involved in creation or execution in these arenas or to *products* that are generated. In either case, the development of these tools requires careful examination of goals, careful delineation of categories of behaviors to be rated or evaluated, and extremely careful selection and training of raters. A machine can score a correctly marked answer sheet, but experts must rate processes and products. The expertise must be twofold: the ability to judge the kind of product or performance and the ability to evaluate the work of students of the particular age group assessed.

Use of product and performance evaluations is becoming much more prevalent. The technology that is being developed for general assessment should be used to improve instruments in the gifted arena. For example, the statewide performance assessment program of Connecticut has included attempts to assess many outcomes that are parallel to those of gifted programs, such as products that reflect real-life solutions to problems or creative solutions to problems.

The Standards Issue

Even when appropriate instruments are found, when change can be assessed over time, and when assessments are deemed reliable, several nagging problems remain. Are the student outcomes that have been achieved better than "nothing," are they better than other gifted programs, or do they meet the standards set for gifted students in a given program or the highest criteria set in the field?

The assumption that the "program for the gifted" is being compared to "no program" is always unfounded. Some educators assume that if there is no formal program for the gifted, there is no program at all. In reality, the question of whether the formal program is "better than nothing" is not a

question because the gifted students will still be in school in *some* educational program even if a formal program does not exist.

So, should the effectiveness of a program for the gifted be compared to other formal programs, to outcomes of the regular curriculum, or to a set of standards or criteria? To what degree is the formal program responsible for meeting the highest standard set by experts in the field of gifted education? To what degree is the formal program *only* responsible for achieving the goals that have been set, even if those goals are inadequate for gifted students and their education? These are values issues that must be examined before undertaking any evaluation process.

Effect of National Achievement Standards on Assessing Gifted Student Achievement

One of the current waves in the reform movement is to set local, state, and national standards for achievement. Many states have already set minimum standards for advancement to high school or for graduation from high school, and several federal officials are proposing national examinations. The most significant negative potential of this movement is that the minimum standards set by such tests will come to be seen as acceptable for all students when, in fact, these standards and examinations fall far short of assessing the predominant goals and objectives for gifted students. The current formats do not even provide reasonable benchmarks that might be used to assess educational achievements of gifted students. Educators involved in gifted programs have a responsibility to become involved in setting standards and to look beyond minimal standards and measures for gifted student assessment.

Changes in Grade Equivalency Scores as Standards

It is tempting to set changes in grade equivalency scores on standardized tests as standards for gifted program assessment. Aside from the validity issues raised above, it is important to recognize that grade equivalency has little meaning for academically gifted students. The concept of grade equivalency is based on differences between the scores of average students across grades. There is no index for assessing reasonable change in grade equivalency scores for gifted students in programs or not in programs over the period of an academic year.

Another faulty assumption is that a given change in grade equivalency is the same regardless of where it occurs on the scale. That is, growth of 1.2 years at one level is considered equal to growth of 1.2 years at another level. In fact, grade equivalency scores are *not* interval scores, so the same change may represent varying degrees of learning or change.

Use of Control Groups to Determine Outcomes

When evaluators have tried to assess the effectiveness of education programs, they traditionally have compared the achievement of students receiving special services with those who are not receiving special services. However, the use of control group comparison in evaluating the effectiveness of services to gifted children has not been accepted because it would mean that some gifted children would not receive services. Educators have been reluctant to restrict access to gifted programs.

Even when control groups are identified, the assessment of program effectiveness is hampered by internal variance within each group. The variability in aptitudes, outside factors affecting the children in the group, teacher instructional competencies, and many other undetermined factors may, in fact, outweigh or interact with the effects of the program.

Solutions to this problem have focused on two strategies: using students as their own controls and using matching control groups. Typically, using students as their own controls involves single subject designs that gather

baseline data on outcome measures, institute instructional intervention and measure outcomes, stop interventions to measure outcomes, and repeat the intervention. This kind of assessment might be used in evaluating models such as the "Revolving Door" component of Renzulli's Schoolwide Enrichment Model. Comparing projects completed while a student is "revolved in" to projects completed as part of regular classroom instruction (using criteria established as goals for the model) demonstrates the differences in types of projects produced while in the special classroom setting and shows the influence of the program on other projects.

Use of Qualitative Evaluation Strategies

The traditional evaluation model assumes that one instructional program is equally good for any gifted student who has been identified. The programming strategies, instructional activities, and evaluation strategies that have been used all assume that one type of program will be equally effective for all gifted students. What may, in fact, be the case is that certain programming strategies and curriculum are effective for certain gifted students having certain characteristics but are not effective for others.

Evaluation designs have not accounted for these effects. Evaluators have not examined the possibility that what is described as the "same" instruction is really quite different from each individual student's perspective and, thus, has quite different effects on individual students in the same program.

It is time to consider using qualitative evaluation strategies in conjunction with quantitative strategies in determining the effects of programs on individual students. Evaluation strategies must provide the opportunity to describe the ways in which gifted students interact with the experiences provided for them, with their teachers, with the way programs are delivered, and with the outcomes they experience. Qualitative studies are needed to determine what works for which individuals under which conditions, what the other intervening factors are that influence success, and how the program deals with those factors.