DOCUMENT RESUME

ED 315 710 CG 022 291

AUTHOR Balow, Irving H.; Schwager, Mahna
TITLE Retention in Grade: A Failed Procedure.
INSTITUTION California Educational Research Cooperative, Riverside.
PUB DATE Feb 90
NOTE 46p.
PUB TYPE Information Analyses (070) -- Viewpoints (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Failure; *Academic Standards; Cost Effectiveness; *Educational Policy; Elementary Secondary Education; *Grade Repetition; Grades (Scholastic); Reliability; Student Promotion; Validity

ABSTRACT

Retention of pupils results in a need for additional teachers, facilities, and materials at a rate approximating the rate of retention. Retention is a more serious problem for the state, which needs to pay most of these increased costs. This paper reviews the research evidence to assess the cost-effectiveness of student retention policies. Following a short review of the history of retention, the paper reviews the literature on the effectiveness of retention, then addresses the issue of retention as a means of maintaining the integrity of the curriculum. It also considers the use of standardized tests or locally developed tests as important elements of promotion standards, and the reliability and validity of letter grades or marks, which provide the professional judgment on which retention may be based. The conclusion is drawn that retention in grade has virtually no benefits for the pupils retained, their classmates, their teachers, or their schools. References are included. (Author/TE)

# RETENTION IN GRADE: A FAILED PROCEDURE

by

**Irving H. Balow**
**Professor of Education**
**Project Investigator**

**and**

**Mahna Schwager**
**CERC Fellow**

**February, 1990**



CALIFORNIA EDUCATIONAL RESEARCH COOPERATIVE

UNIVERSITY OF CALIFORNIA, RIVERSIDE

## THE CALIFORNIA EDUCATIONAL RESEARCH COOPERATIVE

CERC is a unique partnership between county and local school systems and the School of Education at the University of California, Riverside. It is designed to serve as a research and development center for sponsoring county offices of education and local school districts -- combining the professional experience and practical wisdom of practicing professionals with the theoretical interests and research talents of the UCR School of Education faculty.

CERC is organized to pursue six broad goals: These goals serve the needs and interests of cooperating public school members and the University by providing:

- Tangible practical support for school improvement.
- Proven strategies for resolving instructional, management, policy and planning issues facing public education.
- Valuable professional development opportunities for current and future school leaders.
- Support for data based decision-making among school leaders.
- Research, planning and evaluation activities that are meaningfully interpreted and applied to school district problems, and
- Data analysis to assist in generating public support for effective school programs.

In addition to conducting research in these areas, CERC publishes reports and briefs on a variety of educational issues. CERC also sponsors regional workshops for local educational leaders.

---


CALIFORNIA EDUCATIONAL RESEARCH COOPERATIVE

UNIVERSITY OF CALIFORNIA RIVERSIDE

**California Educational Research Cooperative**
School of Education
University of California
Riverside, CA 92521-0128
Phone: (714) 787-3026
FAX: (714) 787-3942

**CERC Executive Staff**

Douglas E. Mitchell
Professor of Education
Director

Jane L. Zykowski
Assoc. Specialist in Ed. Manager

Dena Sowers
Administrative Assistant


---

### CERC MEMBERS

SPONSORING OFFICES OF EDUCATION

**Riverside County Office of Education**

**Office of the County Superintendent of Schools San Bernardino County Office of Education**

SPONSORING SCHOOL DISTRICTS

East San Gabriel Valley ROP
Fontana Unified School District
Hemet Unified School District
Hesperia Unified School District
Jurupa Unified School District
Lake Elsinore Unified School District
Moreno Valley Unified School District
Murrieta Unified School District
Ontario-Montclair School District
Perris Elementary School District
Perris Union High School District
Redlands Unified School District
Rialto Unified School District
Riverside Unified School District
Romoland School District
Temecula Valley Unified School District
Val Verde School District
Victor Elementary School District
Victor Union High School District
Yucaipa-Jonat Unified School District

# RETENTION IN GRADE:
# A FAILED PROCEDURE

A Report Presented to Members of the
California Educational Research Cooperative

by

Irving H. Balow

and

Mahna Schwager, CERC Fellow

February, 1990

## Table of Contents

# SCHOOL FAILURE AND RETENTION: REASONS AND CONSEQUENCES


## Abstract

Though student grade-level retention rates among member districts of California Education Research Cooperative are lower than the estimated national rate (Smith & Shepard, 1987), they are a potentially serious problem. Retention of pupils results in a need for additional school teachers, facilities, and materials at a rate approximating the rate of retention, i.e., a 7 percent retention rate increases expenditures by approximately 7 percent. Retention is a more serious problem for the state because it is the state that needs to pay most of these increased costs. If the increased costs can be justified by the effectiveness of retention in helping the retained pupils, or by its effectiveness in maintaining standards, the integrity of curriculum or instruction, or the maintenance of discipline, then retention should be continued. If it is primarily a burden on the taxpayer and the educational system, it should be abandoned.

This paper reviews the research evidence from these perspectives. Following a short review of the history of retention, the paper reviews the literature on the effectiveness of retention, then addresses the issue of retention as a means of maintaining the integrity of the curriculum. The latter issue includes a consideration of the use of standardized tests or locally developed tests as important elements of promotion standards, and the reliability and validity of the grading or marking of teachers which provide the professional judgment on which retention may be based.

## A Review of the Literature

### Early History

In the mid-nineteenth century, American schools were essentially ungraded. Students moved through the system via content mastery, not incremental age level steps like first, second, third (Beck, Cook, and Kearney, 1960). This was soon to change, however, because of the German influence on American scholars studying in Europe. Scholars were attracted to the graded elementary schools of that country and brought the concept to the U.S. By 1870, every aspect of every school in the country was graded: buildings, teachers, textbooks, curricula, and pupils.

A premise of the graded school was that achievement would be enhanced if the curriculum were graded by year in school if the teacher focused instruction on the curriculum of that grade, and if pupils worked to master that curriculum. However, as soon as gradedness was introduced, it became obvious that some pupils mastered the curriculum with relative ease, and other pupils learned only with difficulty and failed to master any significant portion of the curriculum. The latter group posed a serious problem for the schools: the discipline of the school and the effectiveness of instruction were threatened if pupils were promoted without the necessary skills to succeed at the next level.

Retention in grade, or failure, was introduced as a solution. By 1900, retention in grade was a major problem in education, with the failure rate reaching as high as 50 percent, and with adolescents frequently retained in primary grades. To reduce the impact of a full year of retention, semester, quarterly, and subject retention were tried. With each change, the retention rate became higher (Beck,

Cook, & Kearney, 1960, p. 43).

One rationale underlying non-promotion has the integrity of the school as its focus. It is based on the argument that grade standards signifying definite levels of educational development are needed and that pupils should be required to attain those standards before being promoted. If all students are at or above grade level, teachers can then teach the grade level curriculum to pupils ready to learn it, and avoid contributing to the development of poor attitudes toward school (Beck, Cook, & Kearney, 1960, p. 44). This rationale assumes that:

- 1) all children possess the requisite ability to be successful in school,
- 2) the differences between successive grade levels are quite large,
- 3) the curriculum is appropriate for all children, and
- 4) when a pupil does not master the curriculum, it is the pupil who must be held accountable.

Another rationale for non-promotion is more focused on the pupil. According to this set of beliefs, pupils can be seriously harmed by unearned promotions. Low-achieving students who are promoted to a grade level where they are unable to do the required work suffer emotionally and drop further behind in their school work (Beck, Cook, & Kearney, 1960, p. 44). It was believed that retaining pupils who have not met standards for promotion will provide them with an additional year in which to attain those standards. This rationale assumes:

- 1) that the grade level curriculum is appropriate for all pupils,
- 2) pupils who do not master the curriculum do have the ability to catch up if they are given more time,

- 3) the differences between successive grade levels are quite great, and
- 4) there is greater emotional trauma associated with low achievement at grade level than there is in being placed with a younger age cohort.

An early, intensive examination of some of the assumptions underlying these beliefs was undertaken in 48 Minnesota school systems that had been ranked on the basis of the amount of retention in grade, type of community, SES, size, and teacher qualifications (Cook, 1941b). Nine districts with the highest standards and highest retention rates were compared with the nine districts which utilized "social promotion" or the regular promotion of all pupils in every grade. An analysis of the data revealed that the latter group had higher average achievement at every grade level than did the schools with high standards for promotion. No differences were found by grade level in the variability of achievement in any area of the school, with the range from "high" to "low" being the same in both sets of systems. Further, there was no difference in achievement of students with the same chronological age and same mental age in the two sets of schools. Social promotion neither helped nor hindered achievement, and pupils did not find they could just get by without working.

Publication of the Minnesota study intensified the controversy over school standards and pupil welfare. It has been followed by many other studies using different research methods, asking different questions which sometimes reach different conclusions. This review will attempt to sort out the issues in the controversy and indicate the state of our research knowledge.

## Effectiveness of Retention

**Does retention in grade result in higher achievement and/or better school adjustment than promotion to the next grade?**

A 1975 summary of research on the effects of grade retention identified 159 references, considering this issue (Jackson, 1975). Jackson reviewed all 44 of the research studies which had been completed between 1911 and 1973, paying particular attention to the quality of the studies. He drew two major conclusions:

> There is no reliable body of evidence to indicate that grade retention is more beneficial than grade promotion for students with serious academic or adjustment difficulties . . . . Thus, those educators who retain pupils in a grade do so without valid research evidence to indicate that such treatment will provide greater benefits to students with academic or adjustment difficulties than will promotion to the next grade.
>
> Second, "the accumulated research evidence is so poor that valid inferences cannot be drawn concerning the relative benefits of these two options." (p. 627)

By 1986, the editors of the Phi Delta Kappa publication on promotion and retention noted that the Jackson review was being, "Cited by numerous authors as the most valuable review of literature on the effects of grade retention on the achievement level of low-achieving children" (Barber & Strother, 1976). Despite this lavish praise, Jackson's review was a narrative review, using the reviewer's judgment as the basis for drawing conclusions and synthesizing the effects of a body of literature. When a large and confusing body of literature is available on a topic, this approach is less appropriate than a more quantified approach called meta-analysis (Glass, McGaw, and Smith, 1981).

In meta-analysis, differences between the experimental and control groups are

converted to "effect sizes" by dividing the difference in average achievement between the groups by the standard deviation of the control group. The effect size is a measure of the standardized amount of the difference between the two groups. These effect sizes can then be analyzed using regression analysis or other appropriate techniques.

A meta-analysis review of the promotion/retention literature was carried out by Holmes and Matthews (1984). They analyzed the results of 44 studies of the effects of nonpromotion (including many of those analyzed by Jackson).[1] The results of the meta-analysis showed that retained pupils, when compared to their control groups, achieved .44 SD lower in achievement and that retention had a negative effect on language arts, reading, mathematics, work study skills, social studies and grade point average. In addition, retained pupils were significantly lower than the promoted pupils in social adjustment .27 SD, emotional adjustment .37 SD, and behavior .31 SD, and significantly lower on measures of self-concept and attitudes toward school.

> Those who continue to retain pupils at grade level do so despite cumulative research evidence showing that the potential for negative effects consistently outweighs positive outcomes. Because this cumulative research evidence consistently points to negative effects of nonpromotion, the burden of proof legitimately falls on proponents of retention plans to show there is compelling logic indicating success of their plans when so many other plans have failed. (p. 232)

In another review that found no justification for retention in grade, Smith and Shepard (1987) concluded that retention is a part of the reform program in education that does not work. The evidence shows promotion to be consistently

---

[1]The Holmes and Matthews review included studies published as late as 1981.

better than retention for both achievement and adjustment. Smith & Shepard found retention to be discriminatory to boys, children from poor families, young children, and small children.

While one reviewer who looked only at secondary sources ventures the opinion that the effects of retention seem "murky" (Johnson, 1984), most reviewers have concluded that there is **no good evidence for positive effects** (Jackson, 1975), that retention **does not aid pupil achievement or personal adjustment** (Norton, 1983), or that **retention is harmful** (Holmes and Matthews, 1984; Smith and Shepard, 1987).

In sum, available evidence leaves little doubt that retention is ineffective. Promotion is more effective for increasing achievement and fostering personal, social, psychological, and emotional development.

Nevertheless, retention rates continue to be high according to Smith and Shepard (1987), despite the fact that virtually no evidence recommends the procedure as effective or efficient, and a significant body of research finds the practice expensive to the pupil (Mann, 1986; Shepard & Smith, 1985), the district, the family, and the nation (Nicklason, 1984).

## Recent Studies

Recently, a carefully designed study in the Mesa, Arizona Public Schools provides evidence that under some conditions retention may have more positive results than previously documented (Peterson, DeGracie & Ayabe, 1987). This study tested the effects of nonpromotion when accompanied by stringent

requirements concerning programming for retained pupils. In Mesa, any student being considered for retention has to be identified before the beginning of the second semester and instructional goals established for the pupil for the balance of the year -- goals designed to eliminate pupil deficiencies in achievement. If at the end of the year progress has not been sufficient and the pupil is retained, a specific educational plan for the next year must be constructed. When these policies are followed, pupils who are retained do not just repeat the same experiences a second time. Instead, they complete a program specifically designed to overcome their academic problems.

The Mesa sample consisted of 65 pupils retained in first grade and a matched group of 63 pupils who were promoted, 26 pupils retained in second grade matched with 26 pupils who were promoted, and 15 third grade pupils who were retained matched with 15 who were promoted. The investigators found increased achievement for retained first and second grade pupils, gains that were maintained about two years. By the third year following retention, however, the promoted pupils caught up with the retained pupils in achievement while remaining one year ahead of them in school.

The authors concluded that **retention with remediation** has better results than retention alone, but recommended further research on the question of whether promotion with remediation is ultimately more beneficial. This strategy is supported by Leinhardt (1980) who studied low-achieving kindergarten students promoted to first grade and given a special instructional program. With this support, the promoted students performed at higher levels than retained students

given regular first-grade instruction or those placed in a transition room with special instruction.

Research conducted by Pomplun (1988) in a rural Florida county concluded that retained pupils had higher scores at the end of the next year than the promoted group and that retention was most effective in the lower grades and became increasingly less useful in the higher grades. The conclusions of this study appear to support the common practice of retaining the largest proportion of pupils at the end of kindergarten and a regularly reduced proportion each succeeding year throughout elementary school. However, the study was seriously flawed. The pupils who were retained in grade were compared with matched pupils who had been promoted the year before. Unfortunately, Pomplun used status scores (Normal Curve Equivalent scores) as a basis for that comparison. In effect, he suggests that the average NCE score of 53.38 achieved by retained pupils on the first grade reading test represents a higher level of achievement than the promoted group's average NCE of 39.43 on the second grade test. However, the use of status scores is inappropriate when comparing pupils who are using different levels of the test because they are level specific (Phillips & Clarizio, 1988). Had he used the CTBS scale scores or grade equivalent scores, both of which are developmental scores, he would have found that a first-grade NCE of 53.38 is a scale score of 493 and a grade equivalent score of 1.9. The control group of matched pupils promoted to second grade, on the other hand, had a second-grade NCE score of 39.43 which is a scale score of 539 and a grade equivalent score of 2.2. Therefore, these promoted pupils achieved at a higher level at the end of the next year than did

the retained pupils. Furthermore, at least for the primary and elementary groups, reanalysis of the reading test data shows greater gains from retention at the intermediate grade levels than at the primary grade levels.[2] This contradicts assertions that early grade retention is more beneficial than later grade-level retention.

Retention in grade assumes that an additional opportunity to learn particular content will be beneficial to students who have not attained some minimal level of mastery of that content. This assumption was tested in a recent study of students who failed an eighth grade Minimum Competency Test (Singer, Balow & Ferrett, 1988). These students, usually assigned by district policy to a remedial class, were instead promoted to a regular ninth grade English class then were compared to a group of similar students assigned to the remedial class. Although all students in the study were below grade level (ranging between fifth grade fifth month level and approximately eighth grade fifth month level on a standardized achievement test of reading), those assigned to the regular English class achieved at a higher level than those assigned to the remedial class. That is, taking the higher level class increased the rate of achievement for students below grade at the end of the previous year.

---

[2]Pomplun's study reports data for 22 matched pupils in grades 1 and 2, 15 in 3 and 4, and 10 in secondary, and the article provides no basis for disaggregating the pupils. If all of the primary pupils had been in grade 1 when the study began, the promoted group would have scored .3 of a grade higher than the retained group (2.2 vs. 1.9). However, if all had been in the second grade when the study began, the retained group would have scored .06 grades higher than the promoted group (3.0 vs. 2.94). If one-half were in each grade, the average would have been 2.45 for the retained group, and 2.57 for the promoted group. (These data obfuscates rather than illuminates the effects of retention on achievement.)

Unresolved Problems

Two problems with the research literature on promotion/retention need to be considered. First, no carefully designed experimental studies using random selections of pupils retained and promoted have been conducted. Available studies using pupils matched on low achievement and other measurable characteristics, have a built-in bias in favor of the promoted students. Decisions to promote one low-achieving student while retaining another are generally based on either some kind of evidence or intuition about student potential. Until a substantial study is completed where the decision is first made to retain a group of pupils, then one-half of those pupils are randomly identified for promotion, there will continue to be arguments about the validity of the studies and the efficacy of retention.

The second problem is that arguments about retention/promotion generally ignore the fact that neither action results in dramatic increases in the achievement. When low-achieving pupils are retained, they remain low achievers -- when promoted, they continue to be low achievers. **Neither retention nor promotion is beneficial to the pupils or to the school, if not accompanied by effective programmatic interventions.**

Some evidence found in the literature (Peterson, et al., 1987) suggests that where teachers follow Individual Learning Plans developed specifically for retained pupils, the results are beneficial for the pupils. There is also the suggestion in the literature (Lienhardt, 1980) that if pupils eligible for kindergarten retention are promoted to first grade, but with individually planned first grade programs, their success is enhanced.

## Nonpromotion as a Means of Maintaining Standards

The earliest arguments supporting retention called for the establishment of legitimate standards for promotion for each grade level and the insistence that each pupil be required to attain those standards before being promoted (Beck, et al. 1960, p. 44). The existence of such standards would insure that pupils would try to master the curriculum and that teachers would only have pupils in their classes who had mastered the prerequisite skills for the grade level, resulting in their focusing on grade level material rather than on remedial work. Discipline in the classroom would be enhanced because only serious students would be there, and promotion to the next grade would be important because it would have to be earned.

The research cited in the previous part of this review challenges the assumptions underlying this argument. If grade level retention were an effective means of increasing proper student deportment and compelling pupils to work harder and learn more effectively, the results of retention studies should be much more positive than they are. Moreover, as the research reviewed in this section of the paper reveals -- there is no evidence to support the assumption that student retention practices help to assure the integrity of the curriculum and school standards. The range of individual differences in school ability and school achievement as seen in student scores on nationally standardized tests is large and increases each year a student attends school. Achievement test data from member school districts of the California Educational Research Cooperative, provide insight into the range of achievement that exists at each grade level and the range of

achievement within individual pupils.

## Individual Differences

In a typical classroom students vary greatly in their ability to cope effectively with the curriculum of the school. For example, the incidence of handicapping conditions among school pupils is estimated to be approximately 14%, including: mentally retarded, 2.0 - 3%; learning disabled, 2.0 to 3%; and behaviorally and emotionally disordered, 2.0 - 3.0% (Gearheart & Weishahn, 1980). Most of the students in this group are found in regular classrooms in the public schools. Many are misidentified (Shepard, 1983; Shepard, Smith, & Vojir, 1983), and have very serious learning difficulties that preclude their progressing at a "normal" pace through the school curriculum.

At the other end of the learning continuum are gifted students who comprise about 3% of the school population (Mitchell & Erikson, 1978), and creative students who would not necessarily be evaluated as gifted (Murdoch, 1975; Torrance, 1980). These children tend to learn with ease and to achieve at a higher level than non-gifted or talented children attending higher grade levels.

Other children who have great difficulty meeting the regular standards for grade level achievement are the Limited English Proficient.[3] When they are found in large enough concentrations, they are placed in a bilingual program which may or may not improve their achievement in English (Hakuta & Gould, 1987); when

[3]For an extensive summary of the trends in achievement levels for minorities, see Humphreys (1988).

there are too few for a bilingual class, their classroom instruction is modified through a Bilingual Individualized Learning Plan, much like the Individual Educational Plan for children in special education. This group usually achieves well below grade level standards at least until they have developed a high degree of formal English fluency.

These extreme variations in the ability of students to profit from instruction guarantee that virtually every teacher will confront a wide range of individual differences. Since most learning handicapped children are in regular classrooms even though they may be receiving the help of a resource teacher for a limited period of time each day, teachers are expected to routinely adjust instructional techniques and curriculum content to make it possible for these pupils to grow academically. Similarly, with most gifted and talented pupils in regular classrooms for most of the day, teachers who do not adjust instruction and curriculum to these pupils will encourage boredom and stimulate behavior problems. The Limited English Proficient pupils require a very different set of adjustments. If teachers do not adjust instruction, and curriculum to their limited English conceptual knowledge, the LEP student will be unable to function effectively in the classroom.

While the handicapped, gifted and Limited English Proficient student groups are highly visible and require large scale curriculum and instruction modifications in teaching, there are numerous other individual differences in every classroom that make inflexible standards almost impossible to implement. IQ scores (a gross measure of school ability) in the typical classroom ranges from about 70, for borderline cases of mental retardation, to near genius level 130 or higher.

Although a 70 IQ pupil is considered to be "normal," and a 130 IQ pupil is also normal, we cannot expect them to learn the same curriculum content in the same amount of time. A perusal of the norms manual for the Metropolitan Achievement Tests (MAT), or any other nationally standardized achievement test series, will indicate how unreasonable such an expectation would be in practice.

A Technical Note

To fully discuss the problems presented by large variations in individual student ability, we need to discuss a key issue in student achievement testing. The analysis which follows relies on analyses of grade equivalent scores -- a sometimes controversial basis for comparing student achievement rates. Grade equivalent scores and the scaled scores found on most nationally standardized achievement tests are the most appropriate approach to assessment of overall student development because they have the same meaning regardless of the grade level at which they are earned. These scores also represent real differences when compared across grade levels. Other popular test reporting schemes -- percentile ranks, stanines, and normal curve equivalents -- are status scores; they reveal how a student stands in relation to the grade level group at one point in time. Status scores help obscure the true differences between grade levels because they are developed to differentiate achievement within grade levels, one level at a time. Grade equivalent scores link student test performance to a single performance curve running from beginning kindergarten through the end of twelfth grade. The conclusions supported by grade equivalent score differences are very much the same

as those drawn from a comparison of the performance on a continuous K-12 scale. A strong argument can be made that the grade equivalent score is the most appropriate in assessing educational development in a graded school system (Hoover, 1984).

This technical point can be illustrated by analysis of Otis-Lennon School Ability Test scores. The average score for the Otis-Lennon Test is 100 points at every grade level, and the standard deviation is 15 points. Table 1 shows the scores on that test and the grade equivalent scores on the Metropolitan Achievement Tests (MAT) corresponding to the average for all students, and again for those scoring one and two standard deviations above and below the overall average. As the data in Table 1 reveal, the popular belief that pupils in the normal range of school ability produce similar achievement scores at each grade level is quite at odds with actual student achievement (see Shepard, 1983).

**TABLE 1: School ability test scores and reading achievement grade equivalent norms corresponding to five points on the normative score distribution.**

| School Ability Quotient | 70 | 85 | 100 | 115 | 130 |
|---|---|---|---|---|---|
| Standard Deviations | -2 | -1 | 0 | +1 | +2 |
| Percentile Ranks | 3 | 16 | 50 | 84 | 98 |
| Normal Curve Equivalent | 10 | 29 | 50 | 71 | 93 |
| Grade Equivalent Scores for | | | | | |
| Grade 1 | .8 | 1.3 | 1.7 | 2.3 | 3.1 |
| Grade 2 | 1.4 | 1.8 | 2.7 | 4.4 | 7.2 |
| Grade 3 | 1.7 | 2.5 | 3.7 | 6.3 | 11.1 |
| Grade 4 | 1.9 | 2.8 | 4.6 | 7.9 | 11.5+ |
| Grade 5 | 2.4 | 3.3 | 5.7 | 8.8 | 12.2+ |
| Grade 6 | 2.5 | 3.7 | 6.6 | 11.1 | 12.2+ |

The grade equivalent scores in Table 1 are those corresponding to the third, sixteenth, fiftieth, eighty-fourth, and ninety-eighth percentile ranks in the normative sample selected in 1984. Despite the fact that significant numbers of children were retained at grade level for one or more years, achievement became increasingly divergent with each increase in grade level.

Three percent of all pupils in the norm group were at or below the third

percentile rank and would spend three years in elementary school before attaining the reading achievement level of the average child at the end of the first grade. After six years in school these same students would not achieve as high as the average child does at the end of second grade. At the other end of the spectrum, the highest performing first graders would go directly to the third grade, and would be eligible for a high school diploma before reaching age 11.

Children closer to their class average present similarly complex achievement profiles. After two years in school, for example, the lowest sixteen percent of all students are achieving only at the level of the average child at the end of first grade. After four years they achieve at the level of the average child at the end of second grade. At the end of six years they achieve at the level of the average child at the end of third grade. That is, the lowest sixteen percent of the children in our schools would need to be retained every other year -- taking 12 years to leave sixth grade achieving at the same level as the average student in their class.

Table 2 shows similar data for the Comprehensive Tests of Basic Skills, Form U (CTBS).

**TABLE 2: Comprehensive Tests of Basic Skills, Form U, Corresponding to Five Points on the Normal Score Distribution:**

| Standard Deviations | -2 | -1 | 0 | +1 | +2 |
|---|---|---|---|---|---|
| Percentile Ranks | 3 | 16 | 50 | 84 | 98 |
| Normal Curve Equivalent | 10 | 29 | 50 | 71 | 93 |
| Grade levels | | | Grade Equivalent Scores | | |
| 1 | 1.0 | 1.3 | 1.8 | 2.7 | 3.5 |
| 2 | 1.5 | 1.9 | 2.8 | 4.3 | 6.2 |
| 3 | 1.7 | 2.4 | 3.8 | 5.3 | --- |
| 4 | 2.0 | 2.9 | 4.8 | 7.1 | 10.9 |
| 5 | 2.4 | 3.7 | 5.8 | 9.2 | 10.9 |

The CTBS data show that the below grade level scores corresponding to the 3rd and 16th percentile ranks are very similar to the Metropolitan Achievement Test data in Table 1, suggesting that these levels may be independent of the specific standardized test that the district uses.

## Retention Cannot Work to Control Classroom Variance

Retention policies in many districts stipulate either that no more than two retentions are possible in grades K - 5, or that second retentions are permitted

only under extraordinary circumstances. But the test data just presented demonstrate clearly that -- even with current retention practices -- more than 16 percent of the children in our schools fall more than two grade levels behind in achievement before reaching grade four.

While some districts control the number of times a student can be retained, others establish a standard that all pupils be no more than one year behind grade level on district mandated standardized tests. These policies suggest that for first grade pupils to be promoted to grade 2, they must score 1.0 in reading and/or mathematics; to be promoted to grade 4, third graders should score at 3.0 on the end of year test, etc. Table 3 shows the percentage of pupils who would be subject to retention if this standard were actually applied.

**TABLE 3: The percentage of pupils scoring one year below grade level on end-of-year norms, on two different achievement tests, at five different grade levels.**

Percent scoring at least one-year below grade level:

| Grade Level: | READING | | MATHEMATICS | | TOTAL BATTERY | |
|---|---|---|---|---|---|---|
| | MAT | CTBS | MAT | CTBS | MAT | CTBS |
| 1 | 4 | 4 | 12 | 6 | 8 | - |
| 2 | 29 | 20 | 21 | 11 | 18 | 13 |
| 3 | 34 | 32 | 20 | 18 | 23 | 29 |
| 4 | 41 | 34 | 33 | 25 | 35 | 28 |
| 5 | 41 | 36 | 31 | 31 | 36 | 36 |

With a standard calling for retention of all pupils one or more years below grade level, 4 percent would be eligible to fail by reading score in the first grade -- 36 to 41 percent would be eligible for retention at fifth grade level, depending upon the test used. Six to 12 percent of the pupils would fail by mathematics score in first grade, increasing to 31 percent at fifth grade. As many as 8 percent would fail as a result of the total battery score being at least a year below grade at first grade level, increasing to 36 percent at fifth grade. These figures demonstrate that districts in the California Educational Research Cooperative could not maintain such promotion standards, even when these standards are a part of their policies, because retentions would need to be more frequent than other policies allow. The percent of retentions reported by CERC districts at third, fourth, and fifth grade levels are far below the 20 to 40 percent that would be expected by scores on any district's standardized test.

So far we have been considering normative data rather than data collected in school districts. How accurate are the norms in guiding our expectations when we tabulate real data from real school districts in the Southern California area served by CERC? Table 4 is a tabulation of one district's distribution of achievement across four grade levels.

**TABLE 4: The percentage distribution of total end-of-year battery scores by grade equivalent levels for grades one through four.**

| Battery Grade Equivalent Scores | FIRST GRADE | SECOND GRADE | THIRD GRADE | FOURTH GRADE |
|---|---|---|---|---|
| 12.1 - 12.9 | | | | .6 |
| 11.1 - 12.0 | | | | .6 |
| 10.1 - 11.0 | | | .45 | .6 |
| 9.1 - 10.0 | | | 1.0 | 3.1 |
| 8.1 - 9.0 | | | .45 | .6 |
| 7.1 - 8.0 | | .25 | 2.4 | 5.7 |
| 6.1 - 7.0 | | .25 | 3.8 | 7.9 |
| 5.1 - 6.0 | .3 | 1.6 | 10.8 | 18.6 |
| 4.1 - 5.0 | 1.8 | 1.3 | 12.2 | **18.9** |
| 3.1 - 4.0 | 6.1 | 15.3 | **35.7** | 28.0 |
| 2.1 - 3.0 | 15.8 | **40.5** | 28.0 | 14.8 |
| 1.1 - 2.0 | **66.5** | 40.8 | 5.2 | .6 |
| .1 - 1.0 | 9.5 | | | |
| CLASS AVERAGE: NAT'L PUPIL | 1.9 | 2.5 | 3.9 | 4.8 |
| PERCENTILE RANK: | 61.0 | 41.0 | 54.0 | 52.0 |

The district represented in Table 4 is a high-achieving district, except at grade

2, as represented by the national pupil percentile rank of its mean achievement scores. Nevertheless, 9.5 percent of all first graders are at least one-year below grade level at the end of the year; 40.8 percent of second graders are at least one year below grade level; 33.2 percent of third graders are at least one-year below; and 43.4 percent of fourth graders reach that criterion. Moreover, as the highlighted numbers in Table 4 reveal, the proportion of all children achieving at grade level declines steadily from year to year. Two-thirds of the first graders score between 1.1 and 2.0. Only one child in five scores between 4.1 and 5.0 in the fourth grade.

The overlap in achievement across grade levels is very pronounced, and graphically demonstrates how the differences between grade levels are small in comparison to the differences within grade levels. Although the average difference between any two succeeding grade levels (Table 4) is only about 1.4 grade equivalents, the differences among individuals within a grade level varies from more than five years at the end of first grade to more than ten years at the end of fourth grade. Such differences are virtually impossible to eliminate through administrative groupings.

The individual differences in achievement shown in the above tables are not a new phenomenon in public education, nor are equivalent differences in school ability (Cook, 1948). Indeed, these differences were documented in the earliest studies. The ranges shown in Tables 4 and 5 closely parallel differences found half a century ago by Cook (1941b) who demonstrated that a typical sixth grade class contains approximately an eight year range in achievement in reading and

mathematics.

Some school districts use retention more frequently than others. What is the result of higher levels of retention at the end of eighth grade when students should be ready for high school? Table 5 shows the percentage of eighth grade pupils scoring at various grade levels on a standardized achievement test in one California district that uses scores on this test as one factor in retention decisions.

**TABLE 5: Percentages of students scoring at various points above and below grade level at the end of the eighth grade.**

| PERCENT OF STUDENTS IN: | READING | MATH | BATTERY |
|---|---|---|---|
| More than 3 years above grade level | 1.9 | 1.6 | 1.8 |
| 2 to 3 years above grade level | 4.9 | 4.3 | 3.8 |
| 1 to 2 years above grade level | 9.2 | 16.5 | 9.9 |
| Grade level to 1 year above | 21.8 | 17.8 | 23.2 |
| Grade level to 1 year below | 25.0 | 23.6 | 22.7 |
| 1 to 2 years below grade level | 19.6 | 19.1 | 23.4 |
| 2 to 3 years below grade level | 14.9 | 16.0 | 12.8 |
| More than 3 years below grade level | 2.7 | 1.1 | 2.4 |

Data reported in this table show that at the end of eight to ten years of schooling approximately 36 percent of all pupils are more than one year below grade level on this district test. While this finding is not substantially different than would be expected from the previous data, it does show what happens, or is

likely to happen, when a district uses normative data on a standardized achievement test as a basis for retention standards.

Beyond the inability of the schools to control variation in student achievement, there are technical reasons why achievement tests are inappropriate for controlling grade-to-grade promotion (Airasian and Madaus, 1983). Achievement tests are appropriate instruments for assessing school effectiveness in developing general skills, but they are not valid for assessing acquisition of the specific skills and knowledge set forth in a particular school district's curriculum guides. Contrary to popular belief, nationally standardized tests are of little value in assessing anything other than common elements widely shared by most districts in the country. There are simply too few questions on these tests and too many variations in the scope and sequence of local district curriculum.

Evidence of the insensitivity of standardized tests to the specific curricular concerns of districts is found in a 1985 study of two school systems using four or five different reading series and five different mathematics series (Mehrens & Phillips, 1986; Mehrens & Phillips, 1985; Phillips & Mehrens, 1985; Phillips & Mehrens, 1987). Teachers in grades three and six were rated on the extent to which their in-class curriculum matched the objectives measured by two standardized achievement test series. Measured student achievement was not affected by differences in either text books or teacher emphases on specific district curriculum elements. Two further studies of Mehrens & Phillips support these conclusions. In a series of four studies they concluded that "different textbook series and informal curricula generally have no significant impact on test total,

objective, or item scores."

A final difficulty in using achievement test scores to make grade level retention decisions is the problem of within student achievement differences. Average achievement scores obscure the extent to which individual students do much better in some subject areas than others. Available data indicate that within student variations across subject areas are about 80 percent as large as total between student achievement (Hull, 1927; Cook, 1948). These intra-individual differences are easily demonstrated using the standardized test scores from virtually any classroom. Table 6 shows individual sub-scale scores for the first five pupils, listed alphabetically, for an end-of-year fourth grade classroom in a district using the CTBS, Form U.

**TABLE 6: Intra-individual differences in five fourth-graders as measured by the CTBS, Form U using grade equivalent scores.**

| STU-DENT | READING | | LANGUAGE | | | MATHEMATICS | | |
|---|---|---|---|---|---|---|---|---|
| | VOCAB | COMPRE | SPELL | MECH | EXPR. | COMPU | C&APP | RANGE |
| A | 5.0 | 4.9 | 4.6 | 3.7 | 3.8 | 10.9 | 6.8 | 7.1 |
| B | 3.3 | 3.5 | 2.5 | 4.0 | 4.2 | 5.4 | 4.9 | 2.9 |
| C | 3.7 | 4.0 | 4.6 | 7.4 | 5.7 | 8.2 | 5.2 | 4.5 |
| D | 3.0 | 3.1 | 3.2 | 4.2 | 2.9 | 6.9 | 5.3 | 4.0 |
| E | 10.0 | 8.7 | 6.3 | 6.4 | 7.5 | 5.1 | 6.1 | 4.9 |

The pupil scores shown are representative of within student differences in a typical classroom, and reflect the normal state of affairs in measured achievement at the fourth grade level. The smallest within-individual variability shown in Table

6 is 2.9 years (Student B); the most variability is 7.1 years (Student A). Since this within student variation is common, it is virtually impossible to insist upon grade level standards for all pupils in the three primary curricular areas of reading, mathematics, and language arts. Overall achievement is rather meaningless when specific language or mathematics achievement runs two years ahead or behind the average.

In summary, the normative data for achievement test series commonly used in California, and data from actual practice in California districts demonstrate that retention policies do not, and cannot, maintain the integrity of the school curriculum. Grade level and/or age level groups are just too variable in achievement and in their ability to profit from instruction. Moreover, while we have no direct evidence on the effectiveness of retention on the motivation of pupils, available evidence from studies of self-concept reviewed in the previous section (Holmes & Matthews, 1984) suggest that retention leads to reduced effort.

## Research on Teacher Grading and Testing

Even when district policies specify the use of data, professional judgments by the classroom teachers, the principals and perhaps others in the schools remain a major component in all retention decisions. Such professional judgments are, of course, required in dozens of everyday classroom decisions. Indeed, public schools could not function without them. Nevertheless, retention is an uncommon event with very serious consequences for the pupils, and it is appropriate to examine closely how accurate those judgments may be.

If the grades assigned to pupils are to be accurate reflections of the quality of work they have completed, they must be both valid and reliable. Validity refers to the truthfulness of the assessment, the extent to which the test or the grade measures what it claims to measure. Reliability refers to the consistency of the assessment, the extent to which the test or the grade would remain the same over many trials. Both of these important characteristics of tests, assessments, and grades have been extensively researched.

Reliability

Research on teachers' grading practices suggest that they are often inconsistent and hence unreliable. Typically, for example, teachers do not agree with each other on the grades to be assigned to English papers (Starch and Elliott, 1912), or to mathematics papers (Starch and Elliott, 1913). The situation is so bad that pass-fail decisions for as many as 40 percent of the students depend not on what the students have written, but on who reads the papers (Ashburn, 1936). It has been found that the order in which papers are read may significantly influence the grade assigned to the paper (Stalnaker, 1936; Hales and Tokar, 1975), and that when teachers are asked to re-grade a set of papers that they had graded a month earlier, grades change by as much as 25 points and have only a moderate correlation with the first set of grades (Tiegs, 1952). Compounding this situation, teachers usually do not know the level of reliability of their tests (Stiggins, 1988).

The Minimum Competency Tests (MCTs) California districts are required to develop to determine whether students will be allowed to graduate tend to be more

reliable than the tests individual teachers construct for use in grading students. In studies by Singer and Balow (1987) and by Balow, MacMillan, and Hendrick (1986) the reliability of district developed MCT's varied from a low of .58 to a high of .94. If a test with a reliability of .58 is used just to separate students in the top half of the class from those in the bottom half, more than 20 percent of the students would be placed in the wrong group. If such a test is to be used to assign grades on a typical grading curve, between 40 and 50 percent of all students would get the wrong letter grade (Ebel, 1947).

It has been estimated that the reliability of semester marks or grades assigned by teachers ranges from about .70 to .90, and that their validity is about .70 (Odell, 1950). The reliability estimate means that in a class using "A" to "F" grading and a common grading curve, 23 to 40 percent of the students receive grades inconsistent with their actual achievement (Ebel, 1947).

While several grading practice studies are quite old, there are no newer studies to refute their findings and conclusions. In fact, the Director of the Center for Performance Assessment, Northwest Regional Educational Laboratory, in Portland, Oregon, indicates that the same conditions apply today all over the country and constitute one of the most pressing problems for the improvement of education (Stiggins, 1988).

## Validity

An extensive literature on the validity of teacher grades focuses primarily on evaluation of student essays. Findings from three studies illustrate the findings

from this line of research. First, when teachers are asked to grade essays on content alone, essays with spelling errors or a combination of spelling, grammar, and punctuation errors are assigned lower scores than essays with the same content but without such errors (Scannell and Marshall, 1966). Second, handwriting affects teacher scoring of otherwise identical student essays (Chase, 1968). And third, composition errors predictably lower the grade assigned, even when content is supposed to be the only criterion (Marshall, 1967).

Related research on grading validity indicates that teachers are influenced by a range of extraneous considerations. For example, some research has found that the physical attractiveness of the student and the attractiveness of penmanship had an effect on the grades assigned to essays (Bull and Stevens, 1979), and that even the first name of a student can influence the grade assigned a paper (Erwin and Caler, 1984). Expectations regarding student ability or achievement is another determinant of teacher essay grading (Chase, 1979). Generally, essays with higher reading difficulty tend to receive lower grades (Chase, 1983).

Low validity in essay grade assignment is not limited to public school teachers. At the college level, a similar type of situation exists. Instructors who are extroverts have been found to assign higher grades than instructors who are introverts (Covner, 1982). Grades are not always assigned in accordance with the criteria the instructors claim to be using. Variance in the evaluations of a paper by different teachers, and variance in evaluations of a paper by the same teacher at different points in time shows that teacher judgment is not confined by formal criteria (Emig and Parker, 1976). Harris (1977) studied the grades assigned to

papers by college level teachers who said that their primary grading criteria were content and organization and found that these teachers were, in fact, more concerned with the errors that students made in the mechanics of writing. Unfortunately, the internalized standards of teachers within one discipline and within one school vary so much that assigned grades may relate less to mastery of the content studied and more to the value system of the teacher assigning the grade.

The validity of teachers' marks over the course of a semester or a year has also been studied extensively. On the average, girls are somewhat more likely to get higher marks than boys of equal ability and achievement (Carter, 1952). Students who are well liked by teachers tend to get higher marks than students with equal achievement who are less well liked (Hadley, 1954). Some teachers use high marks as rewards and low marks as punishments for behavior that is not supposed to be related to grading or class achievement (Palmer, 1962). It has been estimated that no more than 49 percent of the variance in students' grades can be accounted for by real differences in their achievement (Odell, 1950).

In a fascinating study examining teachers' thoughts during the grading process, Whitmer (1982) found that the most important teacher judgment variable in assigning grades was the speed with which students moved through the curriculum. However, Burton (1983) found that the level at which one teaches is associated with how one views grading. Primary teachers gave grades because they were required to do so, not because grades were perceived to have any value. Middle school and high school teachers approved of grading and considered grades to be

a fundamental means of communicating with students and their parents. The large majority of these teachers based their grades on tests they administered but did temper measured achievement with professional judgment. Elementary teachers split almost evenly on the use of quantitative data or teacher judgment as a basis for assigning grades. Elementary teachers were more likely to report using non-academic criteria as a basis for grades (Burton, 1983).

A 1928 study documents how performance expectations affect grading (Zillig, 1928). Teachers found fewer of the errors made by students judged to be bright than those made by students judged to be low achievers. In notebooks containing written assignments, the proportion of errors identified by teachers was much higher for dull than for bright students. That is, teachers expect poorer students to do poorer quality work and consequently find more errors. These same teachers expect bright students not to make errors, and do not identify errors even when they are present.

Why such decision-making problems exist is partially explained by a study of randomly selected teachers of third, seventh, and tenth grades in science, social science, and language arts which concluded that teachers do not analyze test results in "the manner espoused and prescribed by measurement specialists." And, "without systematic analysis of these tests, teachers do not have assurance that their tests function as desired" (Gullickson & Ellwein, 1985). At best this means teachers realize less than the full potential of their tests. At worst, many tests may misdirect teachers and their students.

Low validity and reliability of teacher assessments is a central issue in student

assessment. As the Director of the Center for Performance Assessment, Northwest Regional Educational Laboratory notes, while teachers develop many of their own assessments, they are "neither trained nor prepared to face the rigorous demands of classroom assessment" (Stiggins, 1988). While teachers are generally uncomfortable with their own practices, districts rarely provide technical assistance to help teachers in assessment. Moreover, the grading procedures of many teachers are flawed. Teachers lack clear expectations for student performance and use vague criteria for making judgments. The result, Stiggins concludes, are "undependable assessments and inappropriate instructional decisions." In sum, the evidence from research on teachers' tests and grading standards indicates that the data which are used to make serious decisions about students in our schools are far less reliable and valid than they need to be given the fact that each student's future is largely determined by the grades assigned. As a result of the research showing the inadequacies of grading decisions, Dressel (1983) cynically defined a grade as "an inadequate report of an inaccurate judgment by a biased and variable judge of the extent to which a student has attained an undefined level of mastery of an unknown proportion of indefinite material" (p. 23).

## Summary

Evidence collected over many years of research demonstrates unequivocally that retention in grade has virtually no benefits for either the pupils retained, their classmates, their teachers or the schools. Moreover, retention has many disadvantages for the pupils affected. There are also disadvantages for the schools:

retention requires the provision of more school facilities and reduces average achievement in the schools.

There is the possibility that when special programs accommodating the specific needs of retained pupils are implemented, retention may be at least as satisfactory as promotion. Available evidence at this point is not conclusive, however.

One reason grade level retention is an ineffective approach to low student achievement is the rapidly increasing variability in student achievement as grade level increases. Grade to grade differences in the average achievement of normal children are extremely small when compared with the differences in achievement within each grade-level. Standardized achievement test data from today's schools -- with current retention rates estimated at 15 to 19 percent -- show fewer and fewer children operating at or near grade level with each additional year in school. Unless we are prepared for such measures as keeping large numbers of 15 and 16 year old students in the early primary grades and graduating a significant number of 10 to 12 year olds from high school, promotion and retention policy changes have no chance of eliminating multi-year variations in the achievement levels of students in most public school classrooms. Retention policies, at least as they are currently being implemented, retain large numbers of students without producing uniform achievement or maintaining the integrity of grade level curriculum and instruction.

Even though the evidence shows that retention is not effective either as an intervention technique to improve the achievement of retained pupils, as a device for maintaining the integrity of the grade level curriculum and instruction, or as

a mechanism for motivating students to work hard, it appears that parents, teachers, and principals still endorse student retention as a response to low achievement -- especially in the early elementary grades (Byrnes & Yamamoto, 1986).

## BIBLIOGRAPHY


Airasian, P.W., & Madaus, G.F. (1983). Linking testing and instruction: Policy issues. Journal of Educational Measurement, 20, 103-118.

Ashburn, R.R. (1936). An experiment in the essay-type question. Journal of Experimental Education, 7, 1-3.

Balow, I.H. (1964, March). The effects of homogeneous grouping in seventh grade arithmetic. The Arithmetic Teacher, 186-191.

Balow, I.H., MacMillan, D.L., & Hendrick, I.G. (1986). Local option comptenency testing: Psychometric issues with mildly handicapped and educationally marginal students. Learning Disbalities Research, 2, 32-37.

Bangert-Drowns, R.L., Kulik, J.A., & Kulik, C.L. (1986). Effects of frequent classroom testing. Paper presented at the American Educational Research Association, San Francisco.

Barber, L.W. & Strother, D.B. (Eds.). (1986). Student promotion and retention. Hot Topics Series, Phi Delta Kappa, Bloomington, IN.

Beck, R., Cook, W., & Kearney, N. (1960). Curriculum in the modern elementary school. Englewood Cliffs, NJ: Prentice Hall.

Bridgham, R.G. (1972). Ease of grading and enrollments in secondary school science: Pt. 1. A model and its possible tests. & Pt. 2. A test of the model. Journal of Research in Science Teaching, 9, 323-329, 331-343.

Bruce, M.G. (1988). Making the grade or marking time? Phi Delta Kappan, 69 (5), 383-84.

Bull, R. & Stevens, J. (1979). The effects of attractiveness of writer and penmanship on essay grades, Journal of Occupational Psychology, 52, 53-59.

Burton, F. (1983). A study of the letter grade system and its effects on the curriculum. (ERIC 238 143).

Byrnes, D. & Yamamoto, K. (1986). Views on grade repetition. Journal of Research and Development in Education, 20, 14-20.

Carter, R.S. (1952). How invalid are marks assigned by teachers? Journal of Educational Psychology, 43, 218-228.

Chase, C.I. (1968). The impact of some obvious variables on essay-test scores. Journal of Educational Measurement, 5, 315-318.

Chase, C.I. (1979). Impact of achievement expectation and handwriting quality on scoring essay tests. Journal of Educational Measurement, 16, 39-42.

Chase, C.I. (1983). Essay test scores and reading difficulty. Journal of Educational Measurement, 20, 293-297.

Cook, W.W. (1941a). Grouping and promotion in the elementary school (Minneapolis Series on Individualization of Instruction No.2, pp. 26-30). Minneapolis, University of Minnesota.

Cook, W.W. (1941b). Some effects of the maintenance of high standards of promotion. Elementary School Journal, 41, 430-437.

Cook, W.W. (1948). Individual differences and curriculum practice. Journal of Educational Psychology, 39, 140-145.

Covner, T. (1982). The relationship of teacher-personality type to grading freshman composition essays: An empirical study. (ED 245 255)

Dressel, P. (1983). Grades: One more tilt at the windmill. Bulletin. Memphis, TN: Memphis State University, Center for the Study of Higher Education.

Ebel, R.L. (1947, Winter). The frequency of errors in the classification of individuals on the basis of fallible test scores. Educational and Psychological Measurement, 725-734.

Emig, J. & Parker, R.P. (1976). Responding to student writing: Building a theory of the evaluating process. (ED 136 257).

Erwin, P.G. & Caler, A. (1984). The influence of christian name stereotypes on the marking of children's essays. British Journal of Educational Psychology, 54, 223-227.

Fitch, M.L., Drucker, A.J., & Norton, J.A. (1951). Frequent testing as a motivating factor in large lecture classes. Journal of Educational Psychology, 42, 1-20.

Fredericksen, N. (1984). The Real test bias: Influences on teaching and learning. American Psychologist, 39, 193-202.

Gearheart, B.R. & Weishahn, M.W. (1980). The handicapped student in the regular classroom (2nd ed.). St. Louis: Mosby.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). Meta-analysis in Social Research. Beverly Hills, CA: Sage.

Goldberg, L.R. (1965). Grades as motivants. Psychology in the Schools, 2, 17-24.

Gullickson, A.R. & Ellwein, M.C. (1985). Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice. Educational Measurement: Issues and Practice, 4, 15-18.

Hadley, S.T. (1954). A school mark -- Fact or fancy? Educational Administration and Supervision, 40, 305-312.

Hakuta, K. & Gould, L.J. (1987, March). Synthesis of research on bilingual education. Educational Leadership, 38-45.

Hales, L.W., & Tokar, E. (1975). The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. Journal of Educational Measurement, 12, 115-117.

Halpin, G. & Halpin, G. (1982). Experimental investigation of the effect of study and testing on student learning, retention, and ratings of instruction. Journal of Educational Psychology, 72, 32-38.

Harris, W. (1977). "Teacher response to student writing: A study of response patterns of high school english teachers to determine the basis for teacher judgement of student writing." Research in the Teaching of English, 11, 175-185.

Holmes, C.T. & Matthews, K.M. (1984). The effects of nonpromotion on elementary and junior high school pupils: A meta-analysis. Review of Educational Research, 54, 225-236.

Hoover, H.D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GE's. Educational Measurement: Issues and Practice, 3(4), 8-14.

Hull, C.L. (1927). Variability in amount of different traits possessed by the individual, Journal of Educational Psychology, 18, 97-104.

Humphreys, L.G. (1988). Trends in levels of academic achievement of blacks and other minorities. Intelligence, 12, 231-260.

Jackson, G.B. (1975). The research evidence on the effects of grade retention. Review of Educational Research, 45(4), 613-635.

Johnson, J.R. (1984). Synthesis of research on grade retention and social promotion. Educational Leadership, 41(8), 66-68.

Leinhardt, G. (1980). Transition rooms: Promoting maturation or reducing education? Journal of Educational Psychology, 72, 55-61.

Mann, D. (1986). Can we help dropouts: Thinking about the undoable. Teacher's College Record, 87, 307-323.

Marshall, J.C. (1967). Composition errors and essay examination grades re-examined. American Educational Research Journal, 4, 375-385.

Mehrens, W.A., & Phillips, S.E. (1985). Sensitivity of item statistics to curricular validity. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Mehrens, W.A., & Phillips, S.E. (1986). Detecting impacts of curricular differences in achievement test data. Journal of Educational Measurement, 23, 185-196.

Mitchell, P., & Erickson, D. (1978). The education of gifted and talented children: A status report. Exceptional Children, 45(1), 12-16.

Murdoch, J. (1975). The other children: An introduction to exceptionality. New York: Harper & Row.

Nicklason, L.B. (1984). Nonpromotion: A pseudoscientific solution. Psychology in the Schools, 21, 485-499.

Norton, M.S. (1983). It's time to get tough on student promotion -- Or is it? Contemporary Education, 54, 283-286.

Nungester, R.J., & Duchastel, P.C. (1982). Testing versus review: Effects on retention. Journal of Educational Psychology, 74, 18-22.

Odell, C.W. (1950). Marks and marking systems. In Walter S. Monroe (Ed.), The Encyclopedia of Educational Research (pp. 711-717). New York: Macmillan.

Palmer, O. (1962, October). Seven classic ways of grading dishonestly. The English Journal, 464-467.

Peterson, S.E., DeGracie, J.S., & Ayabe, C.R. (1987). A longitudinal study of the effects of retention/promotion on academic achievement. American Educational Research Journal, 24(1), 107-118.

Phillips, S.E., & Clarizio, H.F. (1988). Limitations of standard scores in individual achievement testing. Educational Measurement: Issues and Practice, 7(1), 8-15.

Phillips, S.E., & Mehrens, W.A. (1985). The effects of curricular differences on achievement test data at the item and objective level. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Phillips, S.E., & Mehrens, W.A. (1987). Curricular differences and unidimensionality of achievement test data: An exploratory analysis. Journal of Educational Measurement, 24, 1-16.

Pomplum, M. (1988). Retention: The earlier, the better? The Journal of Educational Research, 81(5), 281-287.

Riley, R.W. (1986). Can we reduce the risk of failure? Phi Delta Kappan, 68, 214-219.

Scannell, D.P., & Marshall, J.C. (1966). The effect of selected composition errors on grades assigned to essay examinations. American Educational Research Journal, 3, 125-130.

Shepard, L. (1983). The role of measurement in educational policy: Lessons from the identification of learning disabilities. Educational Measurement: Issues and Practice, 2, 4-8.

Shepard, L.A., & Smith, M.L. (1985). Boulder Valley kindergarten study: retention practices and retention effects, Boulder, CO: Boulder Valley Public Schools.

Shepard, L.A., & Smith, M.L. (1986). Synthesis of research on school readiness and kindergarten retention. Educational Leadership, 44, 78-86.

Shepard, L.A., Smith, M.L., & Vojir, C. (1983). Characteristics of pupils identified as learning disabled. American Educational Research Journal, 20, 309-332.

Singer, H., & Balow, I.H. (1987). Proficiency assessment and its consequences. California Policy Seminar, Institute of Governmental Studies, University of California, Berkeley.

Singer, H., Balow, I.H., & Ferrett, R.T. (1988). English classes as preparation for minimal competency tests in reading. Journal of Reading, 31, 512-519.

Smith, M.L., & Shepard, L.A. (1987). What doesn't work: Explaining policies of retention in the early grades. Phi Delta Kappan, 69(2), 129-134

Stalnaker, J.M. (1936). The problem of the English examination. Educational Record, 17(41) (Supplement #10).

Starch, D., & Elliott, E.C. (1912). Reliabilty of the grading of high school Work in english. School Review, 20, 442-457.

Starch, D., & Elliott, E.C. (1913). Reliabilty of grading work in history. School Review, 21, 676-681.

Starch, D., & Elliott, E.C. (1913). Reliability of grading work in mathematics. School Review, 21, 254-259.

Stiggins, R.J. (1988). Revitalizing classroom assessment: The highest instructional priority. Phi Delta Kappan, 69, 363-368.

Tiegs, E.W. (1952). Educational diagnosis. (Educational Bulletin No. 18). Monterey, CA: California Test Bureau.

Torrance, E.P. (1980). Creative intelligence and "an agenda for the 80s." Art Education, 33(7), 8-14.

Whitmer, S.P. (1982). A descriptive multimethod study of teacher judgement during the marking process. Paper presented at the annual meeting of the American Educational Research Association, New York.

Zillig, M. (1928). Einstellung und aussage. Zeitschrift fur Psychologie, 106, 58-106.