Journal of Research in Childhood Education
2000, Vol. 14, No. 2

0256-8543/00

Student Achievement of 3rd-Graders in Comparable Single-Age and Multiage Classrooms

Wendy Ong
Phoenix Elementary School District, Phoenix, Arizona
Jeanette Allison
Thomas M. Haladyna
Arizona State University West

Abstract. This study compared reading, writing, and mathematics achievement of students in comparable multiage and single-age classrooms in three different school districts. We sought links between these two types of classroom groupings and 1) gender, 2) Title I and non-Title I status, and 3) Hispanic and non-Hispanic students. We used performance-based standardized achievement tests to assess student abilities of a more complex nature. This study used samples of 3rd-graders from school districts where both single-age and multiage classrooms existed. One threat to the validity of these results was selection bias (Burns & Mason, 1998). This threat was taken into consideration by using a control variable, Title I and non-Title I. We analyzed results by main effects and interaction with this control variable to determine if the type of student taught (e.g., Title I or non-Title I; boy or girl) seemed affected by classroom organization. Results showed higher achievement for regular students in multiage settings when compared with the same population of students in a single-age setting, but no differences were observed for Title I students in both multiage and single-age settings. We concluded that non-Title I students in multiage classrooms were achieving more highly than non-Title I students in comparable traditional single-age settings. However, the hypothesis that multiage grouping might benefit Title I students and other traditionally lower-achieving students was not borne out in this descriptive study.

Advocates of multiage classroom organization claim many psychosocial benefits for students (e.g., Goodlad & Anderson, 1987), but a research synthesis by Veenman (1995a, 1995b) showed few studies with very small effects supporting multiage grouping over single-age grouping. Both Veenman and critics of his review, Mason and Burns (1996), agree about the need for empirical evidence. This evidence should involve studies where the achievement measures represent what the public believes is important to learn.

This descriptive study reports student achievement in comparable single-age and multiage classrooms in three different school districts in an urban county. The purpose of the study was to provide empirical evidence about potential differences in achievement that might occur because of the way students are organized for instruction.

What Is Multiage Grouping?

Single-age classrooms, the dominant classroom organization in the United States, contain students of mostly the same age, who may vary greatly in their academic achievement. Multiage classrooms, in the form of the one-room school house, were prevalent in the 19th century; these classrooms housed students at all ages and developmental levels. With the industrial revolution and mass education, the need to educate all students efficiently led to class organizations consisting of same-age children. An underlying assumption about this type of organization is that students of the same age are developmentally similar.

Experience and research have shown that this assumption is seldom true (Perrone, 1991).

Alternative grouping practices have many labels and constructs, including single-age, multiage, nongraded, and mixed-age, among others (Mason & Stimson, 1996). We deliberately use the term "multiage" in this study. We deem it important to use this term to distinguish it from other terms that frequently do not represent all the elements of a mixed-age classroom organization. Multiage grouping places children who are at least a year apart, age wise and developmentally, into the same classroom (Katz, Evangelou, & Hartman, 1990). Students in multiage classrooms work in small, preferably heterogeneous, groups and they progress academically at their own paces. This type of classroom organization may be more natural of the way children behave in our classrooms. In general, multiage grouping appears to be more aligned with children's natural groupings and learning tendencies, whereas groupings by age are characterized by large variance in achievement and abilities.

Research on Multiage Grouping

Research on alternative classroom organizations has a rich history. One study that particularly influenced the field was a comprehensive review done by Goodlad and Anderson (1959). Subsequent research on cognitive and non-cognitive benefits of multiage grouping has been guided by this publication and its revision (Goodlad & Anderson, 1987).

Non-cognitive. Multiage grouping benefits children in many ways, such as by improving social skills, and reducing discipline problems (Allison & Ong, 1996). Social benefits research is mostly descriptive and qualitative, taking into account participants' testimonies, researchers' detailed field notes, and analyses of video- and audiotaped data. There is no doubt about the growth in social and affective development (e.g., child-child and child-teacher interactions, problem solving, and peer assistance). In particular, data on cooperative problem solving (e.g., Kelley & Fitterer, 1998) reveals noteworthy benefits in favor of multiage classrooms. Although many studies focus on social development, the impact of social gains on academic development, such as literacy, also has important learning implications (e.g., Stone & Christie, 1996a).

A few studies show some negative effects of multiage grouping on older children (e.g., Byrnes, Shuster, & Jones, 1994). In large part, however, research shows advantages for older children (e.g., mentorship, relearning). For example, Stone and Christie (1996a, 1996b) showed that older children's literacy development benefited from more frequent opportunities to expand on basic and comprehension skills while mentoring younger children. By contrast, children in a single-age kindergarten setting showed far fewer helping behaviors. Also evident were fewer social events by which they could expand on literacy learning with peers.

The review by Veenman (1995a) on non-cognitive effects of multiage grouping led to a conclusion that multiage grouping had small, positive effects on self-concept and attitude. This finding, however, might be tempered by Burns and Mason's (1996) hypothesis that these comparison studies likely contain selection biases favoring multiage grouping. Burns and Mason (1998) concluded that, indeed, principals assigned more able students to multiage classes in order to lighten the teaching burden created by such an organization. This assignment made the single-grade classrooms' achievements of single-grade classrooms lower when compared with multi-grade classrooms. Future research needs to address selection bias as a possible threat to validity. Overall, a review of the literature on non-cognitive gains from both qualitative and quantitative research shows multiage organization to have an advantage over single-age organization.

Cognitive. Pavan (1977) reported the first review of studies of student achievement in nongraded classrooms. Research reported then was scant and inconclusive,

although it leaned in favor of nongraded classrooms. Slavin (1987) provided a comprehensive review of research on ability grouping, followed by a larger synthesis of the research on nongraded classrooms by Gutierrez and Slavin (1992). Both reviews offered results and implications favoring alternative grouping. Similarly, Katz, Evangelou, and Hartman (1990) synthesized qualitative and quantitative research and found that mixed-age grouping offered students social and academic gains that are better than, or at least the same as, same-age grouping. Others conducted research reviews with varying results (e.g., Anderson & Pavan, 1993; Veenman, 1995b).

Pavan (1992) reviewed 64 studies on nongraded classrooms, dating from 1968 to 1990, that examined standardized test measures. Within standardized measures, she focused on at-risk students' achievement and mental health. More than half of the studies revealed significant results for achievement and mental health. About 90% of the studies using standardized achievement tests showed nongraded groups doing better or as well as graded groups.

Regarding content area instruction, Stone and Christie (1996a, 1996b) completed a comparative analysis between multiage (kindergarten through grade 2) and kindergarten classrooms, and factored for collaborative literacy learning during sociodramatic play. Children in multiage classrooms displayed marked increases in literacy behaviors compared with their single-age kindergarten peers.

Veenman (1995a) found that no differences existed in the achievement of comparable multiage and single-age students. Using their selection bias theory, Mason and Burns (1996) maintained that a small advantage might exist for single-age teaching. Those authors also bring into play the idea that the quality of instruction in multiage settings may be not as high. They argue that multiage grouping requires a more challenging type of instruction (e.g., curriculum coverage), for which most teachers are unprepared.

Learning Strategies Related to Multiage Grouping

There is increasing evidence on the extent of learning strategies found in multiage classes. These strategies include, for example, peer tutoring and cognitive scaffolding (Brown & Reeve, 1985; Vygotsky, 1978), cooperative learning and heterogeneous grouping (Slavin, 1987), socioemotional versus academic emphasis (Marcon, 1993), and looping (Rasmussen, 1998). Rationales for multiage grouping have been built largely on research resulting from these related practices.

Academic achievement also has been a focus of these studies, and has produced positive results. Advocates of multiage grouping also refer to research from other fields, such as special education and neurology. Commonly, implications from these peripheral fields recommend alternative grouping strategies, such as multiage grouping, to better meet the needs of children with special abilities (e.g., attention deficit disorder with hyperactivity) (Aldridge, Eddowes, & Kuby, 1998). In addition, strong implications from years of brain research have been applied to nontraditional pedagogy (Jensen, 1998; Shore, 1997).

Taken together, these related practices (e.g., peer tutoring, heterogeneous grouping) have built a strong case in favor of nontraditional learning approaches. Results from these alternative practices provide valuable information about constructs within multiage grouping. Even with this type of support, what remains is piecemeal research. What is needed, then, is research on differences in academic achievement in multiage classrooms and single-age classrooms. What is mostly lacking are studies that focus on curriculum-relevant achievement measures in reading, writing, and mathematics.

What Constitutes a Good Measure for Student Achievement?

In most studies of student achievement, the search for a criterion variable leads researchers and evaluators to a common choice: a published standardized achieve-

ment test, such as the Iowa Test of Basic Skills or the Stanford Achievement Test. While these tests are adequate measures of general learning of declarative knowledge, the fact that these tests sample from a large domain of possible tasks limits their usefulness to measure instructional effectiveness over a short period, such as a single year. Also, school district curricula are very explicit about what is and isn't taught in each grade, and the standardized test is hardly a precise instrument to reflect classroom learning for a specific grade and school year. Another limit of these published tests is that they are easily corrupted by teachers, school leaders, and others who will narrow the curriculum and teach to the test in an effort to produce publicly reported test scores attesting to their effectiveness (see Cannell, 1988; Haladyna, Haas, and Allison, 1998; Haladyna, Nolen, & Haas, 1991; Mehrens & Kaminski, 1989; Nolen, Haladyna, & Haas, 1991; Smith, 1991). Because of the way publishers' standardized achievement tests are designed, test scores are unlikely to reflect classroom learning as based on either the district curriculum or the state's content standards. Therefore, studies where publishers' tests are used as a criterion measure are badly flawed, thus casting doubt on the validity of these studies' results.

At the time of the study reported here, Arizona was uniquely engaged in a major reform effort. The state had adopted its own content standards, called the Essential Skills. The Iowa Test of Basic Skills was determined to have a low degree of correlation with these standards (Noggle, 1987), so the state initiated a revolutionary integrated performance assessment in reading, writing, and mathematics, which was designed to reflect the kind of teaching that focused on process and products that required performance aligned to Arizona's content standards. This test sharply contrasted with traditional teaching and testing that called for low-level memorization of fragmented knowledge. In the current study, we decided to focus on reading, writing, and mathematics abilities that were meant to be measured by the integrated performance test known as the ASAP (Arizona Student Assessment Program), and that was linked to the state's Essential Skills.

Objectives of This Study

This study focused on seven questions bearing on the differences between single-age and multiage classrooms as measured by the ASAP. With respect to reading, writing, and mathematics, is there a difference:

1. between comparable single-age and multiage 3rd-grade students?
2. between boys and girls?
3. between Title I and non Title I students?
4. among ethnic groups?

With respect to reading, writing, and mathematics, is there an interaction between classroom organization (single-age and multiage) and

5. gender?
6. Title I and non Title I students?
7. ethnic membership?

Method

Design of the Study

This study employed a quasi-experimental, ex post facto design. In other words, students were not randomly assigned to instructional settings. While such assignment is a very desirable condition in experimental design, according to Mason and Burns (1996) in their criticism of research on multiage and single-age research, there are several reasons why such designs are difficult to implement. First, schools are reluctant to conduct experimental research, especially when theoretical analysis favors one method over another. Second, most universities do not look favorably on experimental studies in which subjects might be exposed to potentially negative treatments.

Therefore, an underlying assumption was that the units of analysis in this study were random with respect to assignment to classes. This assumption is subject to challenge by Burns and Mason (1996), who

suggested that there is evidence for selection bias in studies comparing multiage with single-age classrooms. Their follow-up study (Burns & Mason, 1998) offered substantial evidence for selection bias and the reason for this bias—that principals believe multiage settings are more challenging to teachers, and, therefore, assign more able students to offset the extra pressure put on these teachers.

In this study, we introduced a control variable (Title I/non-Title I) that controlled for the student ability under both classroom organization condition in this study. That is, within each multiage and single-age group, we have two subgroups: one consisting of Title I students who are traditionally lower-level learners, and the other, non-Title I students, who we believe reflect normal and above average learners. We hypothesized differences between Title I and non-Title I students in all three achievement areas tested. We also hypothesized that multiage, Title I, and non-Title I students would outperform their single-age, Title I, and non-Title I counterparts, respectively.

Sample

Three urban Arizona school districts participated in this study. Six schools were selected based on the following criteria: each school had both single-age and multiage classroom organizations at the 3rd-grade level and all students were given the state-based tests in reading, writing, and mathematics. The multiage classrooms had to be extant for at least three years.

The three districts represented rural, suburban, and urban areas of Phoenix's greater Maricopa County. Each district had distinctly different socioeconomic and ethnic composition. District A was a more affluent, rapid growing suburban community of Phoenix, District B was a more established school district with moderate to low socioeconomic status students, and District C was in an adjacent city with a large Hispanic population and low socioeconomic status.

The total sample consisted of 615 3rd-grade students (256 girls and 289 boys). From this sample, 161 students were classified as Title I and 454 students as non Title I. Ethnic membership for this study was determined from the registration of students by parents. The ethnic composition of the sample was 289 non Hispanic and 218 Hispanic students. Because there were few Native American and African American students, they were not considered in this study. Because the gender and minority representation of both multiage and single-age samples were similar, the possibility of selection bias seemed more remote. If a selection bias were prevalent in these three school districts, then a disproportionate number of Hispanic students or Title I students would have been selected for single-age when compared to multiage classrooms. Clearly, this was not the case in this study.

Achievement Tests

As noted earlier, the achievement measures used in this study were integrated performance assessments in reading, writing, and mathematics developed by the Arizona Department of Education for its Arizona Student Assessment Program. An example of a typical item for the personal experience narrative is summarized as follows:

> Students read a personal experience narrative about finding and caring for an injured owl. In the pre-reading activity, students discuss caring for injured animals and birds, and they explore why some birds are called raptors. The students then read the narrative and answer comprehension questions about it. Finally they illustrate and describe their favorite part of the narrative.

Generic rubrics were used by trained evaluators to rate performance.

The developmental history of this assessment is described in a technical report (Riverside Publishing Company, 1994). The Form D test was carefully constructed to reflect the state's content standards, known as the Essential Skills. This integrated assessment shows alpha reliability estimates for reading (.74), writing (.68), and mathematics (.63). While these reliabilities

are low, the consequences of using such measures in a research study jeopardize the power of statistical tests. If statistical significance is not achieved, such results might be attributed to the low levels of reliability of the extant measures used. On the other hand, if statistical significance were achieved, the results can be interpreted as overcoming the limitations imposed by these reliability levels.

With permission from the three school districts, test data were obtained from their archives. No student identification was ever used in this study, in order to protect students' rights and privacy.

Analysis of Data

To answer the seven research questions involving three dependent measures, 4-way analyses of variance (ANOVA) were done for each of the three abilities (reading, writing, and mathematics). Since we had seven research questions, the analysis included four main effects tests (multiage versus single-age, gender, Title I versus non-Title I, and ethnic membership). First-order interactions were limited to questions five, six, and seven. All other interaction effects were grouped with the residual (error or within) in this analysis. With satisfactory statistical power in these analyses, alpha was set at .05 for statistical hypothesis testing. Effect sizes were reported when statistical significance was achieved. These effect sizes were standardized mean differences, using the total group standard deviation for contrasting pairs.

Results and Discussion

Question 1: Multiage and Single-Age Differences

Table 1 contrasts multiage and single-age students for reading, writing, and mathematics. For each ability, multiage students scored higher on the state's integrated performance assessment. The effects were substantial

Table 1
Differences Among Groups

Differences Between Multiage (N=490) and Single-Age (N=190) Students						
Subject	Reading		Writing		Mathematics	
Descriptive Statistics	M	SD	M	SD	M	SD
Multiage	9.5	3.2	4.8	1.6	12.0	4.8
Single-Age	8.6	3.0	4.2	1.6	9.6	4.9
F-ratio and probability	9.9, p<.05		14.8, p<.01		32.4, p<.05	
Effect Size	.294		.363		.488	
Differences Between Girls (N =256) and Boys (N=289)						
Subject	Reading		Writing		Mathematics	
Descriptive Statistics	M	SD	M	SD	M	SD
Girls	9.5	3.2	4.7	1.7	11.0	4.9
Boys	8.1	2.7	4.1	1.6	10.0	5.2
F-ratio and probability	28.2, p<.01		17.6, p<.01		5.2, p<.05	
Effect Size	.456		.364		.203	
Differences Between Title I (N=161) and non-Title I (N= 454) Students						
Subject	Reading		Writing		Mathematics	
Descriptive Statistics	M	SD	M	SD	M	SD
Title I	7.5	2.5	3.9	1.4	9.6	4.9
Non Title I	9.4	3.1	4.6	1.7	12.0	4.8
F-ratio and probability	44.7 p<.01		20.8 p<.01		17.6 p<.01	
Effect Size	.619		.423		.487	
Differences Between Hispanic (N= 289) and non-Hispanic (N=218) Students						
Subject	Reading		Writing		Mathematics	
Descriptive Statistics	M	SD	M	SD	M	SD
Hispanic	8.1	2.8	4.1	1.5	8.9	4.7
Non Hispanic	9.5	3.1	4.8	1.8	12.3	4.4
F-ratio and probability	26.6 p<.01		21.6 p<.01		66.0 p<.01	
Effect Size	.457		.423		.691	

for mathematics but less substantial for reading and writing. The writing assessment was limited to two dimensions for scoring and has a limited range and correspondingly low reliability. Therefore, the effects were not as pronounced as they might have been had the scoring been more comprehensive and reliable.

Question 2: Boy / Girl Differences

Table 1 also summarizes the results of the ANOVAs for boy/girl differences for reading, writing, and mathematics abilities. As shown there, girls outscored boys in all three performance measures. The effect sizes were large for reading, moderate for writing, and small for mathematics. These results are somewhat surprising for mathematics. Boys usually outscored girls on multiple-choice tests of mathematics.

On the performance-based test that involved both reading and writing, however, this advantage for girls reveals a reversal of a long-observed trends. Ryan and Franz (1998) discussed the influence of item format on test performance in various subjects, including mathematics, and Ryan, Franz, Haladyna, and Hammond (1998) provided more evidence of this effect in a statewide assessment that included both elementary and secondary students. Two rivaling hypotheses for the results reported in this study are: 1) that the performance measure more accurately reflects the National Council of Teachers of Mathematics' definition of mathematics ability, which includes the ability to read and understand the problem and express the solution in writing and visually, and 2) that writing and reading ability may contaminate the measurement of mathematics ability. These two rivaling hypotheses have significant implications for how we interpret performance scores in mathematics and deal with boys' and girls' achievement levels.

Question 3: Title I / Non-Title I Differences

Table 1 also provides results for the comparison between Title I and non-Title I students in reading, writing, and mathematics. We expected the non-Title I students to outscore Title I students. Nevertheless, would multiage grouping benefit Title I students more than traditional classrooms would? The results in Table 1 show major differences between the two groups, with non-Title I students outscoring Title I students regardless of classroom organization. Reading scores showed a standardized mean difference of .62, and writing and mathematics had effect sizes of .42 and .49 respectively, in favor of non-Title I students. If differences did not exist, we might reject the test as not reflecting achievement differences that we know, in fact, do exist between these two groups. Another significant finding is that Title I students have a serious reading deficit that detracts from their performance in writing and mathematics. That is, although Title I students may have greater ability in writing and mathematics than that displayed on this test, their low reading ability may lessen their chance to perform highly in these other areas. This result is borne out in the study by Ryan et al. (1998), in which causal modeling was used with statewide samples in Oregon and achievement measures included performance-based tasks.

Question 4: Ethnic Group Differences

Since the demographic composition of the three school districts lacked sufficient representation by Asian American, African American, and Native American students, these groups were not included in the analysis. We made a Hispanic/Non-Hispanic comparison, because many Hispanic students populated these schools. Table 1 shows the results of this analysis. Again, statistically significant differences were noted. The effect sizes were substantial for the differences among groups, with non-Hispanic students outscoring Hispanic students. As with the previous findings involving Title I/non-Title I students, these differences are not surprising given what we know about the performance of students in prior assessments. Another important observation is that the performance of Hispanic students in mathematics is seriously lower than nongraded Hispanics. The ef-

fect size of .69 is 50% greater than the effect size differences in reading and writing. An earlier implication from the Title I/non-Title I contrast was that low reading ability may have contributed to low mathematics performance. This implication may apply more significantly to Hispanic students whose second language is English.

Question 5: Interaction Between Multiage / Single-Age and Gender

No statistically significant interaction was observed between multiage/single-age and gender. Since large effects existed for the multiage versus single-age contrast and for the boy versus girl contrast, interaction effects would be unlikely. This finding suggests the invalidity of thinking that multiage grouping favors boys over girls or girls over boys. The differences observed between boys and girls seem constant within type of classroom organization, namely multiage and single-age.

Question 6: Interaction Between Multiage / Single-Age and Title I / Non-Title I

Significant interactions were detected for reading and mathematics, but not for writing. Table 2 shows the means for the interaction and the effects of logical pairs of groups. Title I students had identical reading scores, whether in multiage or single-age groupings, but a large difference existed between the multiage and single-age non-Title I students. This finding suggests that classroom organization may not affect

Table 2
Interactions Between Multiage / Single-Age and Title I and Non-Title I for Reading and Mathematics

Reading	Title I	Non-Title I	Effect
Multiage	7.5	10.7	.717
Single-Age	7.5	8.9	.456
Effect	.000	.587	
Mathematics	**Title I**	**Non-Title I**	**Effect**
Multiage	9.3	13.7	.894
Single-Age	8.6	9.8	.243
Effect	.142	.793	

Table 3
Interactions Between Multiage / Single-Age and Hispanic / non-Hispanic for Reading and Mathematics

Reading	Title I	Non-Title I	Effect
Multiage	7.5	10.7	.717
Single-Age	7.5	8.9	.456
Effect	.000	.587	
Reading	**Hispanic**	**Non-Hispanic**	**Effect**
Multiage	7.9	10.9	.978
Single-Age	8.1	8.9	.261
Effect	.062	.653	
Mathematics	**Hispanic**	**Non-Hispanic**	**Effect**
Multiage	9.8	15.0	1.057
Single-Age	8.4	11.1	.549
Effect	.284	.793	

achievement for Title I students, but that there may be a sizable advantage for non-Title I students. Multiage Title I students did better on mathematics than their counterparts in the single-age grouping, but the non-Title I students did extremely well in the multiage setting, besting their counterparts in the single-age setting by a very large effect size of .79. The findings in reading and mathematics support an emerging hypothesis that multiage grouping is very effective for non Title I students.

Question 7: Interaction Between Multiage / Single-Age and Hispanic / Non-Hispanic

The final question in this study dealt with the distinction between Hispanic students and non-Hispanic students. As with the findings for Title I/non-Title I students, reading and mathematics results provided a statistically significant interaction, and the interaction for writing was not significant. Table 3 provides descriptive statistics for the two-way interactions.

In reading, multiage and single-age Hispanic students did equally well but considerably lower than non-Hispanic students. In mathematics, both Hispanic and non-Hispanic students in multiage settings scored higher than their single-age counterparts. While these findings reflect the higher achievement for students in multiage groups, these results also suggest that non-Hispanic students seemed to benefit more from multiage grouping than did Hispanic students.

Conclusions and Implications

The first four research questions dealt with differences between groups of students. Multiage students did better than single-age students for reading, writing, and mathematics, but to varying degrees. The differences between boys and girls were very large for reading and writing, but girls also showed a pronounced advantage over boys in mathematics. Differences between Title I and non-Title I students, and between non-Hispanic and Hispanic students, were predictably large, favoring non-Title I and non-Hispanic students. Hispanic students were most notably lower in performance in mathematics than their non-Hispanic peers in both classroom organizations.

The next three research questions dealt with interactions of classroom organization with gender, Title I status, and the Hispanic/non-Hispanic contract. No interaction was observed for writing. Multiage, non-Title I groups scored higher than single-age, non-Title I groups, but this difference was not sustained for Title I single-age and multiage groups. A similar result existed for the Hispanic students in single-age and multiage groups. Students who traditionally score low on cognitive tests (namely, Title I and Hispanic students) did not do especially well, regardless of how they were grouped for instruction.

As noted earlier in this paper, recent reviews of research on multiage and single-age grouping for instruction have led to some controversy (Mason & Burns, 1996; Veenman, 1995a, 1995b). Mason and Burns recommend field experiments of an experimental nature designed to expose subtleties in how these classes are organized and conducted. They also suggest more observational studies that examine the lives of multigrade and single-grade teachers, and appropriate teaching strategies. Veenman (1995a) also recommended observational studies of teachers.

To their recommendations, we add another important component for such studies. The measure of student learning has been hotly contested in recent years. Traditional measures of school achievement have been much criticized (Frederiksen, 1984) as lacking ecological validity. These traditional tests measure declarative knowledge that is foundational to learning more complex behavior. The tests used in this study were clearly developed to reflect learned abilities of a complex nature. The multiage settings have been described by Katz et al. (1990) as having more in common with this shift in emphasis from declarative to procedural knowledge. Thus, it follows that in a state (such as Arizona) undergoing this kind of classroom reform, where teachers are encouraged to develop procedural knowl-

edge, students in multiage settings naturally outperformed students in single-age settings. We believe that future studies should use achievement measures that possess Frederiksen's ecological validity. In other words, the curriculum explicitly underlying instruction should be directly assessed with appropriate tests, and not tangentially focused tests, as we often see with publishers' tests.

Given the appropriate instructional setting, the experimental and observational studies suggested by Veenman (1995a, 1995b) and Burns and Mason (1996) have more potential of exposing differences in achievement if they exist when a more appropriate achievement measure is used.

This study offers encouraging evidence about the benefits of multiage grouping on students' reading, writing, and mathematics abilities. Future studies should focus on achievement measures that better reflect the ecology of classroom teaching, rather than the sterile sampling of low-level behavior provided by standardized achievement tests. In addition, the continued low test performance of Title I students and many Hispanic students, among other minorities, is still reason for concern. If multiage organizations are actually working, then we would expect to see more success in these traditionally lower achieving groups. Using Arizona's content-based tests as measures of student achievement, this study did not uncover any success associated with multiage organization and Title I and Hispanic students. Other factors may be intervening, however. For instance, students with low levels of reading comprehension may be unable to perform on complex performance tests—the lack of strong reading ability depresses performance, for one thing. With increased reading ability, Title I and Hispanic students might perform as well as their counterparts when placed in a multiage classroom. Comprehensive assessments that include reading, writing, and mathematics provide a better means for understanding student's achievement, regardless of how they are organized for instruction, but these assessments are especially needed to tease out achievement in multiage settings.

This study, and others that will follow, sheds more light on the achievement of students in multiage classrooms. More intensive studies of the learning and instructional environments within these comparable classroom organizations might provide more evidence as to why multiage grouping is so widely acclaimed.

Certainly, there is room for improvement in how multiage grouping is investigated. For example, the methods by which some achievement "evidence" has been obtained may be questionable. Mason and Stimson (1996) charged that: "Although many researchers have compared student achievement and affective outcomes in combination and single-graded classes, sound methodological designs are rare" (p. 449). In rebuttal to critiques of multiage grouping, Veenman (1995a) argued that, depending on what one looks for, flaws can be found with virtually any study, especially one that investigated a complex structure such as mixed-age grouping. Also, what is considered "evidence" varies across proponents and skeptics of nontraditional schooling and between quantitative and qualitative researchers.

References

Aldridge, J., Eddowes, E. A., & Kuby, P. (1998). *No easy answers: Helping children with attention and activity level differences.* Olney, MD: Association for Childhood Education International.

Allison, J., & Ong, W. (1996). Advocating and implementing multiage grouping in the primary years. *Dimensions of Early Childhood, 24*(2), 18-24.

Anderson, R. H., & Pavan, B. N. (1993). *Nongradedness: Helping it to happen.* Lancaster, PA: Technomic.

Anderson, R. H. (1993). The return of the nongraded classroom. *Principal, 72*, 9-12.

Riverside Publishing Company. (1994). *Arizona Student Assessment Program Assessment Development Process, Technical Report, Form D1.* Chicago: Author.

Brown, A. L., & Reeve, R. A. (1985). *Bandwidths of competence: The role of supportive contexts in learning and development* (Technical Report No. 336). Champaign, IL: Center for the Study of Reading.

Burns, R. B., & Mason, D. A. (1996). Simply no worse and simply no better may be wrong: A critique of Veenman's conclusion about multigrade classes. *Review of Educational Research, 66*(3), 307-322.

Burns, R. B., & Mason, D. A. (1998). Class formation and composition in elementary schools. *American Educational Research Journal, 35*(4), 739-772.

Byrnes, D. A., Shuster, T., & Jones, M. (1994). Parent and student views of multiage classrooms. *Journal of Research in Childhood Education, 9*, 15-23.

Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average. *Educational Measurement: Issues and Practice, 7*(2), 5-9.

Frederiksen, N. (1984). The real test bias. Influences of testing on teaching and learning. *American Psychologist, 39*, 193-202.

Goodlad, J. I., & Anderson, R. H. (1959). *The nongraded elementary school.* New York: Teachers College Press.

Goodlad, J. L., & Anderson, R. H. (1987). *The nongraded elementary school* (Rev. ed.). New York: Teachers College Press.

Gutierrez, R., & Slavin, R. E. (1992). Achievement effects of the nongraded elementary school: A best evidence synthesis. *Review of Educational Research, 62*, 333-376.

Haladyna, T., Haas, N., & Allison, J. (1998). Continuing tensions in standardized testing. *Childhood Education, 74*, 262-273.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2-7.

Jensen, E. (1998). *Teaching with the brain in mind.* Alexandria, VA: Association for Supervision and Curriculum Development.

Katz, L. W., Evangelou, D., & Hartman, J. A. (1990). *The case for mixed-age grouping in early education.* Washington, DC: National Association for the Education of Young Children.

Kelley, M. F., & Fitterer, H. (April 1998). *Multiage and traditional classroom programs: A comparison of standardized test score data, group cooperation and problem-solving performance.* Paper presented at the Annual Conference of the Association for Childhood Education International, Tampa, FL.

Marcon, R. A. (1993). Socioemotional versus academic emphasis: Impact on kindergartners' development and achievement. *Early Child Development and Care, 96*, 81-91.

Mason, D. A., & Burns, R. B. (1996). "Simply no worse and simply no better" may simply be wrong: A critique of Veenman's conclusion about multigrade classes. *Review of Educational Research, 66*(3), 307-322.

Mason, D. A., & Stimson, J. (1996). Combination and nongraded classes: Definitions and frequency in twelve states. *Elementary School Journal, 96*(4), 339-452.

Noggle, N. L. (October 1987). *Report on the match of the Standardized Tests to the Arizona Essential Skills.* Tempe, AZ: Arizona State University College of Education.

Pavan, B. (1977). The nongraded elementary school: Research on academic achievement and mental health. *Texas Tech Journal of Education, 4*(2), 91-107.

Pavan, B. N. (1992). The benefits of nongraded schools. *Educational Leadership, 50*, 22-25.

Perrone, V. (1991). *Standardized testing.* Urbana, IL: ERIC Clearinghouse on Elementary and Early Childhood Education.

Rasmussen, K. (1998). Looping: Discovering the benefits of multiyear teaching. *Education Update, 40*(2), 1-4.

Ryan, J., & Franz, S. (1998). *Substantive and psychometric relationships among reading, writing, and mathematics achievement with analyses of gender and format-by-gender differences.* Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Ryan, J. M., Franz, S., Haladyna, T. M., & Hammond, D. (1998). *Substantive and psychometric relationships among reading, writing, and mathematics achievement with analyses of gender and format-by-gender differences.* Technical Report 98-02. Phoenix, AZ: College of Education, ASU West.

Slavin, R. (1987). Developmental and motivational perspectives on cooperative learning: A reconciliation. *Child Development, 58*, 1161-1167.

Shore, R. (1997). *Rethinking the brain: New insights into early development.* Washington, DC: National Association for the Education of Young Children.

Stone, S. J., & Christie, J. F. (1996). Collaborative literacy learning during sociodramatic play in a multiage (K-2) primary classroom. *Journal of Research in Childhood Education, 10*, 123-133.

Stone, S., & Christie J. (1996, April). *Collaborative literacy learning during sociodramatic play: A comparative analysis between multiage (K-2) and kindergarten classrooms.* Paper presented at the Annual American Educational Research Association, New York.

Veenman, S. (1995a). Cognitive and noncognitive effects of multigrade and multi-age classes: A best-evidence synthesis. *Review of Educational Research, 65*(4), 319-382.

Veenman, S. (1995b). Effects of multigrade and multi-age classes reconsidered. *Review of Educational Research, 66*(3), 323-340.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.