

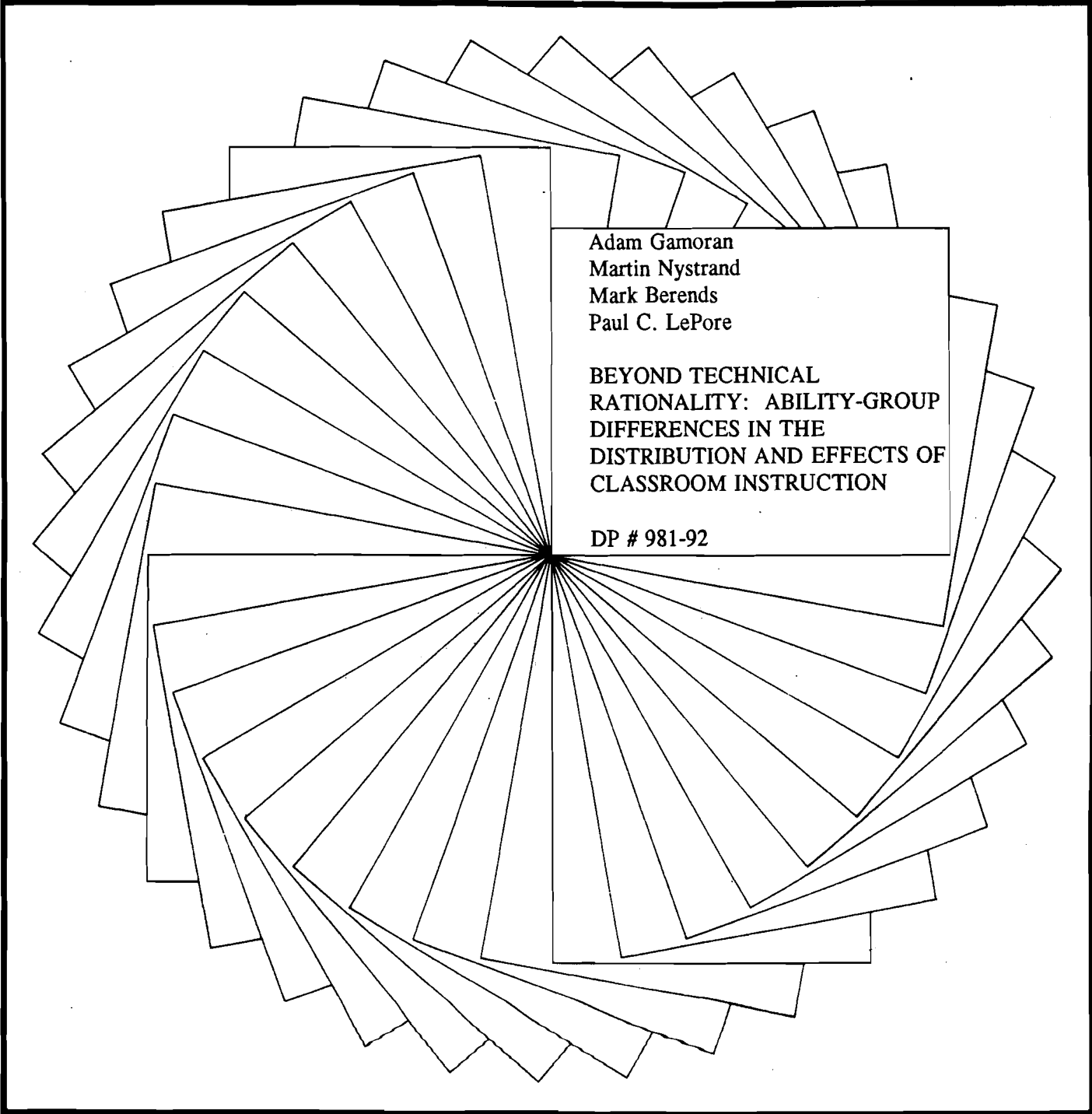
University of Wisconsin-Madison

---

# Institute for Research on Poverty

---

## Discussion Papers



Adam Gamoran  
Martin Nystrand  
Mark Berends  
Paul C. LePore

BEYOND TECHNICAL  
RATIONALITY: ABILITY-GROUP  
DIFFERENCES IN THE  
DISTRIBUTION AND EFFECTS OF  
CLASSROOM INSTRUCTION

DP # 981-92

Institute for Research on Poverty  
Discussion Paper no. 981-92

**Beyond Technical Rationality:  
Ability-Group Differences in the  
Distribution and Effects of Classroom Instruction**

Adam Gamoran  
Department of Sociology  
Institute for Research on Poverty  
University of Wisconsin-Madison

Martin Nystrand  
Department of English  
University of Wisconsin-Madison

Mark Berends  
RAND Corporation  
Washington, D.C.

Paul C. LePore  
Department of Sociology  
Institute for Research on Poverty  
University of Wisconsin-Madison

August 1992

Earlier versions of this paper were presented at seminars at the University of Wisconsin, the University of Chicago, and the Hebrew University of Jerusalem, and the authors appreciate the helpful comments they received at those forums. David Yamane also provided useful suggestions. Research for this paper was supported by a grant to the Institute for Research on Poverty at the University of Wisconsin-Madison from the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. The data were collected under a grant to the National Center on Effective Secondary Schools at the Wisconsin Center for Education Research, University of Wisconsin-Madison, from the Office of Educational Research and Improvement, U.S. Department of Education (Grant No. G-008690007-89). Opinions, findings, and conclusions in the paper are those of the authors and do not necessarily reflect the views of the supporting agencies.

## Abstract

Ability grouping appears to be a rational means of organizing a student body with diverse academic skills. Many observers contend, however, that when students are grouped according to their purported capacities for learning, high-achieving students receive better instruction and increase their achievement advantage over students in other groups. This paper examines the kinds of instruction students receive in honors, regular, and remedial eighth- and ninth-grade English classes. It also assesses the links between instruction and achievement. The authors find that rates of student participation and discussion are higher in honors classes, contributing to the learning gaps between groups. Another finding is that rates of open-ended questions are similar across classes, but that honors students benefit more from such discourse because it more often occurs in the context of sustained study of literature.

**Beyond Technical Rationality:  
Ability-Group Differences in the  
Distribution and Effects of Classroom Instruction**

Ability grouping is the practice of dividing students for instruction according to their purported capacities for learning. It occurs in response to academic diversity in the population of students entering a school. Many educators see ability grouping as "technically rational" in that it provides a structure in which different groups of students can have correspondingly varied instructional experiences. It is supposed to facilitate an efficient matching of instruction to students' needs. Despite this rational foundation, ability grouping has serious shortcomings. It operates contrary to other goals, such as integration, equality of opportunity, and the maximization of individual outcomes. Moreover, ambiguity about the relation between instruction and learning, and a narrow view of instruction generally, have prevented researchers and educators from assessing the efficiency of the system as required by principles of rationality.

With these difficulties in mind, this paper examines the uses and effects of ability grouping in ninety-two eighth- and ninth-grade English classes in eighteen midwest-American secondary schools. Is there evidence of inequitable opportunities across ability groups? If so, does this account for the oft-observed finding of ability-group differences in achievement? What are the connections between ability grouping, instruction, and achievement?

**ABILITY GROUPING AS AN ORGANIZATIONAL RESPONSE TO DIVERSITY**

Despite ambiguity and controversy over the aims of schooling, Americans agree that fostering student learning is an important aspect of the school's mission. In the face of diversity in initial knowledge and skills, and in the capacity to process new knowledge and skills, American schools divide students into categories into which the instructional process, the "technology" of schools, can be correspondingly differentiated. In speaking of instruction as technology, we borrow the language

of organization theorists, for whom technology refers to the materials and activities through which an organization carries out its work to attain goals (e.g., Thompson, 1967; Perrow, 1986). Despite its complexity, instruction is the technology of schools because it is the key mechanism through which learning occurs (Parsons, 1960; Gamoran and Dreeben, 1986).<sup>1</sup> By sorting students for instruction, schools follow a pattern that is common to complex organizations: they create homogeneous subunits to accommodate a heterogeneous input population. As Thompson (1967, p.70) explained:

"Under norms of rationality, organizations facing heterogeneous task environments seek to identify homogeneous segments and establish structural units to deal with each." Structural differentiation under conditions of input heterogeneity is technically rational because it allows for the flexible adjustment of technology to address varying characteristics among the subunits. It occurs not only when the technology is relatively straightforward and routine, such as in a gypsum factory or a steel mill, but also when the technology is more complex and involves feedback, such as in a hospital or a university. In American schools, the differentiated subunits are grades, and ability groups within grades. One would thus expect to find varied technological conditions (different instructional activities and/or materials) in different grades and ability groups.

### Problems of Applying Technical Rationality

If acquisition of knowledge and skills is a central goal of schooling, then the division of students according to capacities relevant for learning is a feature of rational organization. Indicators of capacity may include test scores and judgments of teachers based on experience about students' abilities and willingness to engage in schoolwork. Family background, race, and ethnicity are not characteristics relevant to learning per se, and cannot be justified as the basis for subdivision.

Unfortunately, because of conditions external to the school, when students are sorted on the basis of relevant characteristics they also become divided on the basis of irrelevant characteristics (Oakes,

Gamoran, and Page, 1992). Hence, principles of equity and integration come into conflict with norms of rational organization.

Varied opportunities. Another problem concerns the activities that occur in separate subunits. How are these activities determined? In industrial organizations, differentiated processes are aimed at containing costs and maximizing overall output, and other goals are commonly set aside. Managers of a steel mill, for example, need not care whether they provide "equal opportunities" to different grades of iron ore. Some ore they may discard entirely. Their goal is to produce the most and best steel at the lowest cost. Whether all the ore is "fairly treated" is an absurd question. But this is exactly the issue that schools must confront. Educators desire not only to produce high levels of learning overall, but to maximize the learning of each individual, and to minimize differences among individuals. Sometimes educators make greater investments in subunits of lower initial capacity, even when the payoff is below average. For example, special education programs are among the most costly, and remedial classes are often smaller than other classes. This strategy would be irrational for a manufacturing firm, but it makes sense in the school because it addresses goals of individual maximization and equity.

More typically, it seems, school staff invest more in their higher-performing groups and classes. Although this tactic may be technically rational (that is, it invests resources where they are seen as most likely to pay off), it works contrary to equity and to the success of many individual students. Dar and Resh (1986), for example, argued that sorting students by ability creates a resource-rich environment for high-group students and deprivation for students in low groups, because the intellectual capacities of classmates constitute important classroom resources. In many schools, moreover, teachers with the best reputations are assigned to honors classes, and less-experienced and/or less-successful teachers are relegated to remedial classes (Finley, 1984; Talbert, 1990). Other researchers have reported that secondary school teachers spend less time preparing and are less

enthusiastic with low-track classes (Rosenbaum, 1976; Vanfossen, Jones, and Spade, 1987).

Instruction in low-track classes tends to be more fragmented, dwelling on isolated bits of information, and it progresses at a slower pace (Oakes, 1985; Page, 1991). Even in elementary schools, observers have seen less-conducive learning environments in low-ability groups within classes (Allington, 1980; Eder, 1981). Differential instruction has often meant lower-quality instruction for low-status groups, and several reviewers have speculated that such instructional differentiation has resulted in widening achievement gaps over time (Gamoran and Berends, 1987; Murphy and Hallinger, 1989).

Ambiguous technology. In light of such speculation, it may seem surprising that little research has been done that actually measures the links between ability grouping, instruction, and achievement (Oakes, Gamoran, and Page, 1992). The absence of such evidence is particularly problematic from the standpoint of technical rationality, for it means that educators lack information that would allow them to adjust their grouping systems to improve performance. Technical rationality implies a capacity to assess the value of technology and modify it for greater efficiency and effectiveness. In the case of ability grouping, there is little evidence that the practice leads to higher levels of learning overall, and debate over whether it leads to greater inequality (Slavin, 1987, 1990; Gamoran and Mare, 1989). Yet it is difficult to interpret these findings without information about what occurs in the different groups and tracks and how students' varied experiences contribute to variation in learning. In organizational terms, it is hard to judge the adequacy of the structure without information on the performance of the technology.

One reason for insufficient research is that instruction is an ambiguous technology: cause-effect relations are poorly understood, and there is no consensus on what constitutes effective instruction, nor on how to measure instructional variation. Teachers, after all, are not merely applying treatments to objects, but are interacting dialogically with students, who are not inert raw material, but sentient, intentional subjects (Nystrand, 1992). Teachers have the dominant role, but

precisely what action they should take is far from clear, in part because their efforts are sensitive to the characteristics and reactions of students (Jackson, 1968). Hence, it has been difficult to postulate a cause-effect chain to describe the relation between teaching and learning. Because of this ambiguity, most researchers and educators have turned away from instruction (technology) to focus on the institutionally legitimized categories (structure) of education, such as grade levels, diplomas, and certification (Meyer and Rowan, 1978). As Meyer (1980) explained, schools (and researchers) pay little attention to what gets taught and how, but much more attention to who gets taught by whom. Research on ability grouping has followed this pattern: Much work has examined where students are assigned and what achievement they obtain, but as Slavin (1987, 1990) discovered, few studies of ability grouping and achievement have also considered what students experience inside their classes.

Narrow view of instruction. Where instruction has been measured, it has been narrowly conceived. Most views of instruction construe learning as the result of what teachers do, that is, what they plan and provide for students. In this conception, teaching is the one-way transmission of knowledge from teacher and texts to students, and the results are typically tested by examining students' recall of this information. This notion is inconsistent with the insights of cognitive psychology stressing the nature of learning as an active, constructive process (Piaget, 1937). If knowledge is partly the result of what the knower brings to the learning situation, and not simply a duplication of what someone else says, then conceptions of instruction must somehow accommodate this insight. Narrow views of instruction also emphasize recall at the expense of higher-order thinking, and privilege coverage over depth (Newmann et al., 1988). A more comprehensive view of instruction must cast a wider net if it is to capture the import of significant instructional activity.

The narrow view of instruction glosses over the role that writing, reading, and classroom talk--that is, instructional discourse--play in the formulation of the knowledge and information that is taught and learned. Instructional discourse potentially does far more than serve as a mere conduit for



"packaging" and transmitting knowledge and information; it is not a neutral vehicle of communication. To the contrary, writing, reading, and classroom talk fundamentally shape the knowledge and information that is taught, and some sorts do a far better job than others. As writing-across-the-curriculum programs have discovered, the language of the classroom can be instructionally significant when it helps students sort through, develop, and reflect on their thinking (Nystrand, 1977; Langer and Applebee, 1987).

Adding to knowledge about ability grouping requires enhancing our understanding of instruction and improving our capacity to measure it. Just as managers consider the interface of structure, technology, and output to assess the efficiency of their firms, we need to address the links between grouping, instruction, and achievement to understand how learning occurs. This effort is especially challenging in schools because the technology of instruction is complex and ambiguous. To the extent that we can measure the distribution and effects of instruction across ability groups, we may learn whether and how the educational system affects conflicting preferences (i.e., the desire for efficient administration on the one hand, and the desire for integration and equity on the other).

#### Adding to Knowledge about the Effects of Ability Grouping

Empirical research cannot determine which among competing preferences should receive highest priority, but it can show us how our different educational goals are affected by current secondary-school practice. This paper addresses a key problem: How does instruction vary across ability groups, and how does that variation affect student achievement? The problem has three main components: (1) Is the quality of instruction higher in honors classes and lower in remedial classes? (2) Does a given amount or quality of instruction have higher payoff in high-status classes? Advocates of ability grouping maintain that differential instruction benefits low-group as well as high-group students, whereas critics argue that low-group students are poorly served. (3) To what extent does differential instruction lead to inequality of outcomes among ability groups? If instructional

variation leads to achievement inequality, we may consider what changes in instructional practices could lead to greater equality.

---

### Adding to Knowledge about the Effects of Instruction

This study is feasible only if we have some way of measuring the quality of instruction. We conceptualize instruction more broadly than the narrow view of lecture, recitation, and coverage through which teachers inform their students. Rather than defining instruction as what teachers "do to students," we define it in terms of how teachers and students interact; and to measure the extent of their interaction, we focus on the quality of their instructional discourse.

The most obvious features of high-quality instructional discourse are high student participation and correspondingly low offtask behavior, which both result, of course, when students and teachers interact extensively. Though important, high student participation is not a sufficient measure of high-quality instructional discourse, because the level of student activity alone indicates little about the nature of student engagement. Students are procedurally engaged when they pay attention, do their assignments, and consistently conform to the requirements of school tasks. Students who are procedurally engaged, however, are not necessarily intellectually engaged in the issues and content of their studies. For this reason, we need to examine variation across classes in the substantive quality of teacher-student discourse.

In addition to student participation, one of the most important features of high-quality instructional discourse is its coherence (Nystrand and Gamoran, 1991b). Some teachers carefully frame lessons and activities in terms of previous lessons and activities, and, as a result, classroom talk frequently refers to previous classroom talk. More than this, these teachers also have their students discuss what they have read, write about what they have read, read and discuss before writing, and so forth. During question-and-answer exchanges, effective teachers also follow up on student responses by incorporating previous student answers into subsequent questions in a process linguists call uptake

(Cazden, 1988; Collins, 1982, 1986). This continuous interweaving of writing, reading, and talk helps students relate topics of instruction, and reinforces and builds upon previous learning.

Some teachers also increase the coherence of their instruction by asking questions that build upon students' concerns and interests, and, as a result, skillfully help students relate these concerns to the content of instruction and learning. For example, instead of asking only test questions (questions which are characteristic of recitation and which teachers ask when they are looking for particular answers), skillful teachers also ask authentic questions, or questions for which the teacher avoids prespecifying answers (e.g., How did you like the chapter you read last night? Did the story end the way you expected? Who do you think was the most important character?). Authentic questions are effective because they promote student ownership and help students coherently relate the new information of instruction to what they already know and/or have experienced. Authentic questions are also important because they signal to students the teacher's interest in what students think, as well as the importance the teacher attaches to thinking and not just remembering (Nystrand and Gamoran, 1988). Another aspect of high-quality discourse is discussion, the free exchange of opinions and information among teachers and students, without continual prompting by questions from the teacher.

### Ability Grouping and Instruction

Prior research suggests that the quality of discourse is higher in high-track classes and lower in low tracks. Procedurally, more offtask behavior occurs in low-track classes, teachers spend more time on discipline and less time on instruction, and students spend less time on homework (Oakes, 1985). Substantively, instruction in low-track classes is more often fragmented, emphasizing isolated bits of information instead of sustained inquiry (Page, 1991). In a pilot study (Gamoran, 1989; Nystrand and Gamoran, 1988), we found that students in low-ability eighth- and ninth-grade English classes answered true-false, multiple choice, and fill-in-the-blanks questions four to five times as frequently as did their high-group counterparts. In responding to the papers of students in low-ability

classes, teachers commented twice as much about spelling, 1.8 times as much about punctuation, and twice as much about grammar. In their responses to high-track students' papers, by contrast, teachers commented nearly twice as much about content. And although they met with students in both low- and high-track classes about as infrequently in writing conferences (about once a month on average), they discussed spelling 2.6 times as much with students in low-track classes in these conferences, and content 1.9 times as frequently with high-track students.

Not only is there reason to believe that effective instruction occurs more often in high-ability classes, but such instruction may be most important just where it occurs least. Scholars who write about at-risk students emphasize the need to promote ownership and meaningfulness in schoolwork to counteract the alienation that is common for such students (Wehlage et al., 1989; Wehlage and Smith, in press). To the extent that authentic questions and discussion serve these ends, their positive impact may be greater in low-ability classes in comparison to high-ability classes, where students may be more motivated by external rewards such as grades (Newmann, in press).

In addition, student misbehavior occurs and is treated differently in high and low tracks. As Metz (1978) observed, when high-track students disengage from schoolwork, they do so in a way that still allows them to carry out the task at hand. Passing notes, reading unrelated books, and making humorous remarks occur in the context of making it through the school day, while still getting one's schoolwork done. Thus, disruptive behavior in honors classes is less likely to impede students from carrying out their work, in comparison to regular and especially low-track classes where offtask behavior is part of students' rejection of classwork. Moreover, teachers react differently to misbehavior in high-ability classes. According to Metz (1978), students who are loud or speak out of turn may be seen as overeager but worth engaging in honors classes, whereas similar behavior generates reprimands in low-track classes. For these reasons, procedural disengagement may impede learning more in low-ability than in high-ability classes.

Our aim in the present study is to assess variation in the quality of instructional discourse across tracks, and to discern the impact of instructional variation on student achievement. This analysis will show how ability grouping works, and will help us judge the ways in which it succeeds and fails as a technically rational procedure.

## DATA AND METHODS

The sample for this paper comes from a two-year study of twenty-five secondary schools. The schools were located in nine communities in the American midwest, including rural, urban, and suburban areas, and public and Catholic schools. Overall about four English classes per school participated in the study, but this varied by the size of the school: fifty-eight eighth-grade classes were distributed among sixteen middle and junior high schools, and fifty-four ninth-grade classes were studied the following year in nine high schools for which the middle schools served as feeders. In smaller schools, all classes participated, and in larger schools, classes were selected to represent the different ability-group levels defined by the school (e.g., honors or accelerated, regular, and basic or remedial). About 90 percent of students in the selected classes participated in the study.<sup>2</sup>

The analysis is restricted to ninety-two high, regular, and low classes in ten junior high/middle schools and eight high schools. Heterogeneous classes were excluded from the present analysis for three reasons: (1) In the ninth-grade study, heterogeneous classes were used only in a small, rural school, and a school-within-a-school in an urban school, so homogeneous/heterogeneous differences were confounded with school differences; (2) Standardized test scores, which serve as "ability" measures to help control for preexisting differences among students from different tracks, did not exist for most of the eighth-grade heterogeneous classes; (3) The main issue for this paper is not the difference between homogeneous and heterogeneous classes, but the differences in the distribution and effects of instruction among the homogeneous classes. The eighteen schools

remaining in the analysis included two urban high schools and their three feeder junior highs in an ethnically diverse, mainly working-class area; one urban high school in a less diverse, more middle- to upper-middle-class locale; one suburban school and two feeder middle schools in an upper-middle-class community; two small-town/rural schools, each with one high school and one junior high (thus, four schools in all); and two Catholic high schools with three feeder K-8 schools, which served urban and suburban, predominantly middle- and upper-middle-class white students.

We visited each class four times, focusing mainly on the time spent in different activities and on the questions asked by teachers and students (see below). Students took tests and filled out questionnaires in the fall and spring, and teachers also filled out questionnaires in the spring. Of 1968 students who began the year in the ninety-two classes, 1750 (89 percent) participated in the study in the fall and spring. Listwise deletion of student-level missing data reduced the analysis sample to 1564 students (89 percent of study participants, 79 percent of the total).

#### Background and Achievement Data

We measured learning with a year-end test of literature achievement. Because we assessed instruction as the quality of instructional discourse, we designed a test that required students to engage in discourse about the material they had covered during the year. The chances for detecting the effects of schooling are greater if one tests students on what they were actually taught, rather than on a standardized body of information (Walker and Schaffarzick, 1974).

The test posed a series of questions about the novels, short stories, and plays that were assigned during the year. For each class, we selected five readings that had been covered, choosing items that were representative of the overall curricula. The questions ranged from simple recall ("Describe the ending of the story") to ones requiring in-depth understanding ("Relate the conflict of the story to its ending and theme"). The questions were the same for each class, but the stories differed, depending on what had been covered during the year. Each test was scored by two trained

readers on dimensions such as extent of recall, depth of understanding, understanding of characters' motivations, and so on. When the scores differed by more than one point on any given dimension, the test was given a second reading. Scores from the two readers were averaged, and inter-rater reliability was calculated as correlations of .90 in the eighth-grade sample and .82 in the ninth-grade group.

Prior reading and writing skills. We administered two tests at the beginning of the year to account for differences among students in reading and writing skills. One was a multiple-choice test of reading comprehension, based on National Assessment of Educational Progress (NAEP) items. The eighth and ninth graders read different stories, but the results were calibrated on similar scales. This test also included a brief writing sample. The second test consisted of a fifteen-minute essay, for which eighth graders were asked to write about a person or event that was important to them, and the ninth graders wrote about a special place or possession. This test was scored by two readers on level of abstraction (Britton et al., 1975) and coherence of argumentation (Applebee, Langer, and Mullis, 1985), and the marks were summed across dimensions and averaged across readers. The inter-rater correlation was .70.

"Ability." From school records, we obtained data on student performance on standardized tests administered by the districts. We recorded national percentile scores, which we transformed to normal curve equivalents. Unfortunately, the districts employed several different instruments, and while most were administered in the spring of the previous year, some were given in the previous fall, a full year before our arrival. This would not matter much if all the scores were truly normed to the national population, but the extent to which that is the case is unknown. To account for measurement error introduced by the standardized tests, we used the scores not as distinct variables, but as indicators of a common underlying trait, which we termed "ability." For each student, ability was indicated by a math score and a reading comprehension score. The measurement model for this

latent variable yielded reliability estimates of about .54 for the math score and .44 for the reading score. These values are lower than is typical for such tests, presumably due to differences across districts.<sup>3</sup>

We used the ability measure despite its problems because of the danger that the effects of ability grouping could be inflated by unmeasured differences among students assigned to the different groups. Slavin (1990) has argued that all observed effects of grouping in correlational studies are likely due to such selection bias. While selection bias can never be completely ruled out in the absence of random assignment, the present study offers a more rigorous set of controls than has been used in nearly all comparable studies. In research on high school tracking with another rich data set, Gamoran and Mare (1989) found that using a similar set of controls eliminated the correlation between unobserved selection factors and outcomes. If selection bias is still present in the current study, it is likely to be very small.

Other background variables. Further controls for student background differences were indicated by dummy variables for sex (1 = female, 0 = male) and minority status (black or Hispanic = 1, others = 0). Last, student socioeconomic status was indicated by an unweighted linear composite of father's education, mother's education, the higher in status of father's or mother's occupation, and the availability of a list of home resources. These background data were drawn from student questionnaires. Means and standard deviations of all variables are listed in Table 1.

#### Indicators of Ability-Group Positions

Recent writers have criticized survey studies of grouping and tracking for using ambiguous indicators of track positions (Gamoran, 1989; Lucas, 1990; Lucas and Gamoran, 1991). U.S. studies of national data typically rely on student self-reports of whether their programs are best described as academic, general, or vocational. Although this indicator is useful when tracking is viewed as a social-psychological construct (see Gamoran, 1987), its value as a structural indicator is limited.



TABLE 1

Means and Standard Deviations<sup>a</sup> of Variables: Eighth- and Ninth-Grade Ability-Grouped English Classes

Variable	All Classes	Honors Classes	Regular Classes	Remedial Classes	Source of Data
<b>Dependent Variable</b>					
Literature achievement	15.822 (6.776)	19.774 (6.110)	15.625 (5.933)	9.838 (5.268)	Researcher-administered
<b>Background Variables</b>					
Sex (female = 1)	0.504 (0.500)	0.512 (0.500)	0.503 (0.500)	0.491 (0.501)	Student questionnaire
Minority (black or Hispanic = 1)	0.187 (0.390)	0.096 (0.295)	0.154 (0.361)	0.426 (0.495)	Student questionnaire
SES	0.001 (0.815)	0.372 (0.738)	-0.071 (0.768)	-0.415 (0.808)	Student questionnaire
Fall reading score	27.418 (7.630)	29.913 (7.185)	28.594 (6.430)	20.097 (6.906)	Researcher-administered
Fall writing score	5.995 (1.390)	6.767 (1.366)	5.714 (1.247)	5.488 (1.274)	Researcher-administered
Standardized math score <sup>b</sup>	64.237 (19.240)	80.412 (14.462)	60.828 (16.074)	46.847 (13.258)	School records
Standardized reading score <sup>b</sup>	62.110 (17.771)	77.662 (13.344)	59.167 (13.559)	44.476 (13.032)	School records
<b>Instructional Variables</b>					
Percent of reading completed	83.700 (23.612)	86.971 (19.433)	83.101 (24.704)	79.938 (26.104)	Student questionnaire
Percent of writing completed	87.387 (20.364)	92.112 (15.904)	86.475 (20.753)	82.076 (23.948)	Student questionnaire
Percent offtask	4.202 (5.666)	2.220 (2.673)	3.794 (3.854)	8.583 (9.718)	Classroom observation
Percent of authentic questions	20.554 (16.900)	16.991 (12.125)	24.234 (18.849)	16.404 (15.839)	Classroom observation
Percent of questions with uptake	19.967 (11.566)	20.553 (8.440)	21.656 (12.399)	14.396 (11.999)	Classroom observation
Minutes of discussion time	0.563 (1.409)	1.163 (2.041)	0.167 (0.323)	0.653 (1.582)	Classroom observation
Discourse coherence <sup>c</sup>	11.970 (6.514)	11.016 (6.570)	13.109 (6.464)	10.436 (5.989)	Teacher questionnaire
Number of students	1564	480	793	291	

Note: All schools are in the American midwest.

<sup>a</sup> Standard deviations are in parentheses.

<sup>b</sup> Normal curve equivalent of national percentile.

<sup>c</sup> In scale of times per week (see appendix for questionnaire items).

This is not so much because students may be incorrect; other data sources also carry the danger of unreliability (Gamoran and Berends, 1987). Rather, the ambiguity of the survey indicator stems from an underlying assumption: that virtually all schools are in fact divided into academic, general, and vocational programs. Yet recent observers report that such programmatic tracking has waned, at least in formal terms (Oakes, 1985; Moore and Davenport, 1988). Instead, students in both junior and senior high schools tend to be stratified by performance on a subject-by-subject basis.

Whereas a student's track position (e.g., academic or general) is often ambiguous, there is little disagreement about the ranking of courses within a particular subject. In the case of English, the great majority of secondary schools distinguish among levels such as honors or accelerated, regular or average, and basic or remedial (Oakes, 1985; Moore and Davenport, 1988). For this study, English classes are categorized as honors (including classes labeled high, advanced, and accelerated), regular, and remedial (including classes termed low and basic). These categories were unambiguously described by school staff. Student membership in particular classes was taken from class rosters and was verified by classroom teachers. The sample was not large enough to distinguish among schools having two, three, or four ability levels, but the grouping systems were similar across schools in that all students were assigned to particular English classes based on how well they do in English, rather than on how well they do in all subjects overall (see Slavin [1987] on the importance of this similarity). In four cases, teachers divided their time between two ability groups in a single room.

### Measures of Instruction

For this study we have relied on seven key indicators of instructional discourse. More indicators were available in the data, but we narrowed our focus on the basis of preliminary exploratory factor analyses, inspection of reliabilities in confirmatory factor analyses, and the theoretical centrality of particular indicators (Gamoran, Berends, and Nystrand, 1990; Nystrand and

Gamoran, 1991a). In an early analysis of the eighth-grade data, we had used a composite indicator of discourse quality (Gamoran and Nystrand, 1990), but we learned subsequently that there was no single underlying factor that incorporated the diverse measures (Gamoran, Berends, and Nystrand, 1990). Consequently, we now use the seven variables as indicators of distinct aspects of discourse quality.

We obtained three measures of student participation. Two came from the spring student questionnaire: students' reported frequency of completing their writing and reading assignments. (See the appendix for the wording of questionnaire items.) The other came from classroom observations: the percentage of students visibly offtask during question-answer sessions.

For discourse coherence, we used a composite of six teacher-questionnaire items that asked about the interconnections among different classroom activities: the extent to which teachers asked students to (a) write about what they read, (b) discuss their writing before and after the writing is done, (c) discuss readings, (d) relate readings to other readings, (e) relate discussions to previous discussions, and (f) discuss what other students have written about (see appendix).

Uptake was computed as the percentage of questions that followed up on what someone had said previously, averaged over the four observations. In the following exchange, for example, the teacher's second question uses uptake:

Teacher: Why did Atticus need Aunt Alexandra at this time?  
 Student: To keep Scout safe.  
 Teacher: Why would Scout be safe with Aunt Alexandra?

In this dialogue, which occurred during a discussion of To Kill a Mockingbird in a ninth-grade class, the teacher had specific answers in mind. Even though she was asking the student to draw conclusions rather than simply recite the story, these questions test students' knowledge and use of information instead of encouraging them to construct new ideas, and we refer to them as "test" questions. In contrast, "authentic" questions treat students' ideas as legitimate knowledge in their

own right. A bit later in the lesson, this teacher asked students to speculate about alternative paths the story might have taken: "What are some ideas for Atticus not having Aunt Alexandra come?" Here, she was asking an authentic question, showing interest in students' ideas rather than testing for a prespecified answer. We computed the percentage of teacher questions that were authentic, averaged across the four observations, as another indicator of discourse quality.

Finally, we counted the number of minutes per day devoted to discussion. Discussion is defined more narrowly than simply teacher-student discourse; it refers to the free exchange of information among teachers and students, without the usual question-response-evaluation structure of ordinary recitation (Mehan, 1979). Often during discussion, students speak to one another without interruption by the teacher (Nystrand and Gamoran, 1991b). We focused on discussion because from the standpoint of instructional discourse, it is qualitatively different than other classroom activities which are heavily dominated by teachers.

### Statistical Models

Our initial questions are descriptive. They concern the compositions of the different groups and the differences among groups in the quality of instructional discourse. Subsequently, we turn to the analytic questions of net achievement differences between groups, the effects of instruction on achievement, and the extent to which variation in the distribution and effects of instruction produce ability-group differences in achievement.

To address the analytic issues, we used maximum likelihood methods. We chose this approach because it permitted us to specify the latent "ability" construct described above. It also provided tests for the comparative fits of various alternative model specifications. We divided the data into the three groups: honors, regular, and remedial classes. First, we set aside the instructional data and estimated models of ability-group differences in achievement net of the exogenous variables (sex, minority status, SES, fall reading and writing performance, and ability). After selecting a

baseline model and estimating achievement differences between groups, we added the instructional data to the model. We compared the fits of models in which the instructional variables were constrained to be the same across ability groups with models that permitted different effects in different groups. After selecting the best-fitting model, we reexamined the achievement gaps between groups under various instructional circumstances.

## RESULTS

Does ability grouping curtail economic and social integration in secondary schools? Previous writers have maintained that minority students and economically disadvantaged students are overrepresented in low-status groups and tracks (e.g., Oakes, 1991), and our data conform to that pattern. As Table 1 shows, whereas the sample as a whole consisted of nearly 20 percent minority students, honors classes had just half that proportion while remedial eighth- and ninth-grade English classes averaged more than twice the total sample mean. The contrast is even greater if we focus on the district in our sample with the highest proportion of minority students. In this district, located in a working-class urban area, 52 percent of the students were black or Hispanic, but the proportion minority was 26 percent in the honors classes, 52 percent in regular classes, and 65 percent in remedial classes. Similar patterns appear for the social class composition of the ability groups: In the total sample, honors classes averaged .37 standardized units above the mean in SES while low-ability classes stood at .42 standardized units below the mean.

These findings are far from new, and they cannot be used as evidence that assignment procedures were discriminatory. As previous studies have shown, the direct impact of sociodemographic conditions on track assignment is small, compared to the overwhelming importance of academic performance (e.g., Gamoran and Mare, 1989; Gamoran, in press). The point here is to

show that addressing one goal--reducing academic diversity within instructional groups--conflicts with another goal--ethnic and economic integration within schools.

### Distribution of Instruction among Ability Groups

Are there inequities among ability groups in the conditions of instruction? Table 2 displays class-level means of instructional variables for the different types of classes. As expected, students in honors classes exhibit the most consistent participation, and students in remedial classes are least engaged in their schoolwork. These findings replicate those of other studies, both in their consistent patterns and in that the differences, while statistically significant, are not large substantively (Oakes, 1985; Gamoran and Berends, 1987).

In contrast, most aspects of instructional discourse did not differ significantly between class types. If anything, regular classes contained higher proportions of authentic questions and questions with uptake, as well as a higher degree of coherence, but these differences are not statistically significant. Only discussion favored honors classes over other classes, but it was not a common occurrence even there, averaging only about 75 seconds per day. Authenticity and uptake were also infrequent, with less than a quarter of questions having one or both of these qualities in regular classes, and smaller proportions elsewhere.

The results are consistent with descriptions of classroom discourse as dominated by teachers and emphasizing reproduction rather than production of knowledge (Mehan, 1979; Goodlad, 1984). We did not find evidence of especially fragmented and recitation-oriented instruction in low-ability classes. Although we observed significantly more discussion in high-ability classes, it remains to be seen whether this difference is related to achievement gaps in light of its infrequency. Similarly, even though participation was greater in honors classes and lower in remedial classes, it is not yet clear whether these differences help account for achievement gaps between ability groups. An additional

TABLE 2

**Class-Level Means of Instructional Variables: Eighth- and Ninth-Grade  
Ability-Grouped English Classes**

Instructional Variable	Class Type		
	Honors	Regular	Remedial
<b>Participation</b>			
Percent of reading completed <sup>a</sup>	87.791	81.986	80.417
Percent of writing completed <sup>a</sup>	91.306	84.657	82.546
Offtask in class <sup>a</sup>	2.043	4.079	6.840
<b>Discourse</b>			
Percent authentic teacher questions	16.635	22.943	16.975
Percent of questions with uptake	19.409	21.315	17.059
Minutes of discussion per day <sup>a</sup>	1.224	.200	.643
Coherence of instruction	10.865	13.351	10.367
Number of classes	24	44	24

**Source:** Student questionnaires, teacher questionnaires, and classroom observations, in eighteen schools in the American midwest.

<sup>a</sup> F-test for differences between class types is significant at  $p < .05$ .

possibility is that differences between groups in the effects of instruction, rather than differences in instructional means, produce differential achievement.

### Effects of Ability Grouping on Achievement

Before bringing together grouping, instruction, and learning, we need first to determine whether students in different types of classes obtained varied achievement, net of preexisting conditions. To address this question, we estimated a model in which the effects of all background variables were constrained to be equal across honors, regular, and remedial classes. This model fit the data reasonably well, with a chi-square of 55.89 and 32 degrees of freedom.<sup>4</sup> Table 3 shows that each of the background conditions contributes significantly to literature achievement. Girls, whites, and high-SES students scored higher than boys, minority students, and the economically disadvantaged, respectively. Students with higher initial test scores and higher estimated ability also performed better on our test at the end of the year.

Since all the effects were constrained to be equal, the only differences among models for the three groups are in the intercepts. Consequently, the intercepts reveal differences in achievement between groups, net of background conditions. These show gaps of .843 points between the honors and regular classes, and another 1.147 points between the regular and low-ability classes. These differences are not large--they constitute about 12 percent and 17 percent, respectively, of the total-sample standard deviation--but because they occurred within a single academic year, they need to be taken seriously.<sup>5</sup>

To test for statistical significance of ability-group differences in achievement, we cannot compare the intercepts with their standard errors; that tests whether the intercepts differ from zero, and we need to test whether they differ from each other. This question is addressed by comparing this model to another in which the intercepts are constrained to be equal across groups. We estimated the equal-intercept model and found that its fit was significantly poorer, yielding a chi-square of 66.62



TABLE 3

**Maximum Likelihood Estimates of Background Effects on Literature Achievement  
in Eighth- and Ninth-Grade Ability-Grouped English Classes  
(N=1564 students)**

Independent Variables	Effect	Standard Error
<b>Background</b>		
Sex (female= 1)	1.051**	.263
Minority (black or Hispanic= 1)	-1.292**	.347
SES	.661**	.179
Fall reading score	.292***	.020
Fall writing score	.629***	.108
Ability	.101***	.018
<b>Intercepts</b>		
Honors classes	-1.707	1.337
Regular classes	-2.550*	1.081
Remedial classes	-3.697**	.930

**Source:** Authors' calculations based on student questionnaires and reading and writing tests, in eighteen schools in the American midwest.

**Note:** Chi-square equals 55.89 with 32 degrees of freedom.

\* Coefficient is twice its standard error.

\*\* Coefficient is three times its standard error.

\*\*\* Coefficient is four times its standard error.

with 34 degrees of freedom. The chi-square difference between these two nested models is 10.73, with 2 degrees of freedom, a difference that is significant at  $p < .01$ . Hence, we conclude that the type of class students attended made a small but significant difference in their achievement.

#### Ability Grouping, Instruction, and Achievement

To what extent were the achievement differences produced by variation in the distribution and effects of instruction? We first included the instructional variables using the same specification as we used for the background variables--that is, no differences between class types in the effects of instruction on achievement.<sup>6</sup> Fit statistics for this model are presented in the first row of Table 4. This model fit the data fairly well, but we had reason to question the assumption of equal instructional effects across groups. Our conceptual formulation, and some preliminary analyses, suggested that offtask behavior, authenticity, and discussion might have varied effects, and we estimated this model next. As shown in the second row of Table 4, this model fit significantly better. Subsequent modifications, however, failed to improve the fit. Hence, the data suggest that completion of reading and writing, coherence, and uptake exert similar effects in honors, regular, and remedial classes, but the effects of offtask, authenticity, and discussion vary by class type.

Table 5 displays the results of the best-fitting model. Each of the variables with similar effects across groups contributes positively to achievement: students who report completing more of their reading and writing scored higher, as did those in classes with more uptake and more coherence among instructional activities. The effects of the other instructional variables are more complex: Offtask behavior led to lower achievement in regular and remedial classes, but not in honors classes; authentic questions resulted in higher achievement in honors classes but lower achievement in remedial classes; and discussion benefited honors students but reduced achievement for those in regular classes. The effects of the participation variables appear especially remarkable; for example, a 10 percent increase in offtask behavior would reduce achievement by about 1.25 to nearly 2.0

TABLE 4

## Alternative Models of the Effects of Background and Instruction on Achievement in Eighth- and Ninth-Grade English Classes

Model	Chi-Square	Degrees of Freedom	Comparison to Previous Model		
			Chi-Square Difference	Degrees of Freedom	P
(1) Same effects of instruction in each ability group	123.69	67			
(2) Varied effects of offtask, authenticity, and discussion	86.33	61	37.36	6	< .01
(3) Model (2) plus varied effects of uptake	86.19	59	0.14	2	> .50
(4) Model (3) plus varied effects of writing completed	85.74	57	0.45	2	> .50
(5) Model (4) plus varied effects of reading completed	85.65	55	1.09	2	> .50
(6) Model (5) plus varied effects of coherence	85.17	53	0.48	2	> .50

Source: Authors' calculations based on student questionnaires, teacher questionnaires, and classroom observations, in eighteen schools in the American midwest.

TABLE 5

**Maximum Likelihood Estimates of Background and Instructional Effects on Literature Achievement in Eighth- and Ninth-Grade Ability-Grouped English Classes  
(N=1564 students)**

Independent Variables	Effect	Standard Error
<b>Background</b>		
Sex (female= 1)	1.188***	.252
Minority (black or Hispanic= 1)	-.652	.339
SES	.155	.174
Fall reading score	.202***	.024
Fall writing score	.512***	.103
Ability	.121***	.018
<b>Instruction</b>		
Completion of reading	.022**	.006
Completion of writing	.025**	.007
Offtask in class		
Honors classes	.149	.092
Regular classes	-.193***	.044
Remedial classes	-.124***	.028
Authentic teacher questions		
Honors classes	.056*	.022
Regular classes	.003	.010
Remedial classes	-.050*	.017
Uptake	.063***	.013
Discussion		
Honors classes	.277*	.129
Regular classes	-1.510*	.591
Remedial classes	.045	.169
Discourse coherence	.158***	.022
<b>Intercepts</b>		
Honors classes	-8.502***	1.385
Regular classes	-7.081***	1.207
Remedial classes	-7.144***	1.061

**Source:** Authors' calculations based on student questionnaires, teacher questionnaires, reading and writing tests, and classroom observations, in eighteen schools in the American midwest.

**Note:** Chi-square equals 86.33 with 61 degrees of freedom.

\* Coefficient is twice its standard error.

\*\* Coefficient is three times its standard error.

\*\*\* Coefficient is four times its standard error.

points in the remedial and regular classes, a loss of almost 20 percent to 30 percent of a standard deviation. Similarly, students who did half their reading and writing assignments would score more than 2.0 points lower than those who completed all their work, other things being equal. Effects of the discourse variables are generally more modest, but large enough to be substantively as well as statistically meaningful.

What does this mean for the effects of ability grouping on achievement? As Table 5 shows, completion of reading and writing assignments contributes to literature achievement, but does it contribute to achievement gaps between honors, regular, and remedial students? It does, since a significantly higher percentage of honors students than regular students, and a significantly higher percentage of regular students than remedial students, complete their reading and writing assignments (see Table 2). Uptake and coherence, while having a significant positive effect on literature achievement, do not contribute to the gap in achievement between the three groups of students, since the differences between class types in the mean values of these variables are not significant.

Authentic questions, discussion, and offtask behavior--unlike completion of reading and writing, uptake, and coherence--do not have the same effects on literature achievement for each class type. Therefore, whether or not they contribute to gaps in achievement depends on how frequently these three occur in honors, regular, and remedial classes. At the very least, they demonstrate that the same instructional quality can result in unequal achievement in the different types of classes. The greater the incidence of each, the wider the achievement gaps. For example, if there were no authenticity, discussion, or offtask behavior, the intercepts would capture all of the differences in achievement, suggesting little difference between regular and low groups ( $-7.081 - -7.144 = .063$ ) and higher achievement in regular than high classes ( $-8.502 - -7.081 = -1.421$ ). At low levels, say 15 percent authenticity, half a minute per day in discussion, and no offtask, achievement would be roughly similar across all class types. More realistically, when instructional conditions are at the

averages for all classes, achievement is similar in regular and low classes but about two points higher in high-ability classes. Hence, achievement gaps result from a combination of differences in the levels and the effects of instructional conditions.

### Interpreting the Varied Effects

How should we interpret the differences across groups in the effects of offtask behavior, authenticity, and discussion? Only the first was anticipated: We predicted that offtask behavior might be more harmful in lower-status classes because there, such activity more often reflects resistance to schooling, while in honors classes misbehavior does not necessarily indicate rejection of schoolwork. In addition, high-group students who are not themselves offtask may be less distracted by offtask behavior than students in regular and remedial classes. In light of the positive (though nonsignificant) coefficient for offtask in high groups, another interpretation must be considered: Offtask behavior may occur in honors classes after students have mastered the material. Perhaps students who have figured out the answers and completed their work are afterwards more likely to relax and misbehave. In that case, high achievement may lead to offtask behavior, rather than the reverse. This interpretation challenges the causal ordering specified in our model.

The differential effects of authenticity and discussion did not conform to expectations. To understand the pattern in the results, we returned to the data to examine the content of authentic questions and the contexts of discussions in different classes. We discovered that teachers in honors English classes were much more likely to ask authentic questions about literature, whereas authentic questions in remedial English classes pertained to a wide variety of topics. One teacher in a remedial class, for example, asked authentic questions about test-taking: "How do most of you feel about test-taking?" Another example was brainstorming: "What things would you associate with lying in the sun?" By contrast, authentic questions in high-ability English classes generally focused on ideas and issues found in literary texts. Overall, we counted 73.4 percent of authentic questions had to do with

literature texts in honors English classes, but only 31.3 percent of authentic questions in remedial English classes concerned the texts students were reading.

The pattern for discussion was similar but the interpretation is less clear-cut. Almost all the instances of discussion in honors English classes concerned literature, whereas only half the discussions in remedial English classes were about texts students were reading. However, two-thirds of the discussions in regular English classes were on literary texts. Thus, we are not able to account for the substantial negative effect of discussion in the regular classes.

Like the results for offtask, the pattern for authenticity could also be interpreted as reflecting a mis-specified causal sequence. This interpretation would suggest that authentic questions about topics other than literature are the teacher's response to, rather than a contributor to, low-track students' poor performance in literature. At present, we are unable to test among these competing causal chains. More generally, we cannot test whether high-quality instruction produces higher achievement, or higher achievement leads to better instruction. Our model does not presume a causal order between participation and discourse variables--we assume these conditions are interrelated--but on the basis of our controls for prior ability and achievement, we have assumed that instructional conditions affect year-end achievement rather than the reverse.

#### CONCLUSIONS: BEYOND TECHNICAL RATIONALITY

The results of this study cast doubt on the utility of judging ability grouping according to standards of technical rationality. In part, this conclusion reflects the inherent tensions that result from conflicting preferences. The system of grouping students by performance clearly works against efforts to promote economic and ethnic integration. In addition, it tends to divide students who are less engaged in their schoolwork from those who participate more fully, creating a more disruptive environment where it is most harmful.

At the same time, the role of ability grouping in magnifying achievement inequality also reflects the way ability grouping is typically implemented. We observed more consistent participation and more discussion time in honors classes than in other types of classes. Although we did not find ability-group differences in the quality of discourse on other indicators, we discovered that the content of discourse differs dramatically, with more attention to literature in honors English classes than in regular or remedial classes.

On the one hand, one might argue that the types of authentic questions and discussions that occurred in low-ability classes were just what was called for. Perhaps by holding brainstorming sessions, talking about test-taking, and so on, these classes were meeting students' needs. Although these encounters failed to improve (and perhaps impeded) literature achievement, they may well have contributed in other areas. On the other hand, this conclusion admits defeat in the effort to engage low-achieving students in serious academic work. Students in remedial classes were not denied access to authentic questions, but they had far fewer opportunities to address such questions in the context of literature, one of the major foci of secondary school English. Hence, it was not the interactive style but the content of the interaction that favored honors over regular and remedial classes.

To the extent that ability grouping continues to be used, analyses such as this one can contribute by showing what needs to be done to improve its outcomes. The data suggest that inequality could be reduced by raising the caliber of both instructional content and instructional discourse in regular and remedial classes. According to our results, this would make it possible for students outside the honors level to benefit from high-quality discourse.



**Appendix**

## Wording of Questionnaire Items

Student questionnaire

Completion of reading: "About how often do you complete your reading assignments for this class?"

Completion of writing: "About how often do you complete your writing assignments for this class?"

Response categories: Never, almost never, less than half the time, about half the time, more than half the time, most of the time, every time. Responses in these categories were scored 0, 10, 33, 50, 67, 90, 100, respectively.

Teacher questionnaire

Discourse coherence:

"About how often do students in your class write about (or in response to) things they have read?"

"About how often do you discuss writing topics with your students before asking them to write?"

"About how often do you and your class discuss the readings you assign?"

"When you ask students about their reading assignments in class, how frequently do you . . . ask them to relate what they have read to their other readings?"

"About how often does your class relate its discussion to previous discussions you have had?"

"About how often do you and your class discuss things students have written about?"

Response categories: Never, less than once a month, once a month, two to three times a month, once a week, more than once a week, every day. Responses in these categories were scored on a monthly scale of 0, .5, 1, 2.5, 4, 10, 20. Then they were summed across items, and divided by 4 to convert to a times-per-week scale.

## Notes

<sup>1</sup>Conceiving of instruction as technology does not necessarily imply a narrow transmission of knowledge from teacher to students, for as Thompson (1967), among others, has shown, client-serving organizations tend to employ more complex technologies involving feedback. Note also that "technology" here does not merely refer to electronic aids to instruction, as the term is used in the educational vernacular.

<sup>2</sup>About half the cases in the analysis are students who were included in the study twice, once as eighth graders and a second time in ninth grade. These students are represented twice in the data set. Although this may artificially increase the correlations among the predictors to some degree, the increase does not appear serious. The eighth- and ninth-grade data were obtained in separate years, and measures of classroom instruction, the key independent variables, were completely independent from one year to the next. We could find no meaningful differences between students who participated once and those who participated twice, and we gain much statistical power by pooling the data across grades. We also tested for differences between grades in the effects of the background variables on achievement, and found no significant differences.

<sup>3</sup>The schools were less successful than we were at obtaining data from all students; about 15 percent of students for whom we had complete data on background and prior achievement lacked standardized test results. Scores for these students were imputed on a district-by-district basis from the background and prior achievement data.

<sup>4</sup>The fit of the model could be improved slightly by allowing all background variables to have different effects across groups (chi-square difference = 24.62, d. f. difference = 12, difference is significant at  $p = .016$ ). We chose to estimate the more constrained model because (a) we had no strong prior grounds for predicting between-group differences in effects of background variables; (b) the relaxed model would greatly complicate the estimation of track effects; and (c) after instructional

variables are added, the improved fit from allowing varied background effects is not statistically significant (chi-square difference = 19.01, d.f. difference = 12,  $p = .088$ ).

<sup>5</sup>We compared these results to an ordinary least squares (OLS) regression in which the standardized test scores were included as single-item variables. This analysis yielded track effects that were considerably larger, at close to 1.5 points for each gap. Hence, our model is a more conservative test for track effects, and probably does a better job of accounting for preexisting differences among students assigned to different types of classes, compared to an OLS analysis.

<sup>6</sup>To simplify the model, we did not specify causal paths from the background variables to the instructional indicators, but left these relations as zero-order correlations. This specification does not affect the estimation of direct effects of background and instruction on achievement.

## References

- Allington, R. (1980). Teacher interruption behaviors during primary-grade oral reading. Journal of Educational Psychology, *72*, 371-374.
- Applebee, A., Langer, J., and Mullis, I. (1985). Writing: Trends across the decade, 1974-84. Princeton: Educational Testing Service.
- Britton, J., Burgess, T., Martin, N. McLeod, A. and Rosen, H. (1975). The development of writing abilities, 11-18. London: Macmillan.
- Cazden, C. (1988). Classroom discourse: The language of teaching and learning. Portsmouth, N.H.: Heinemann.
- Collins, J. (1982). Discourse style, classroom interaction, and differential treatment. Journal of Reading Behavior, *14*, 429-437.
- Collins, J. (1986). Differential instruction in reading. In J. Cook-Gumperz (ed.), The social construction of literacy. Cambridge: Cambridge University Press.
- Dar, Y., and Resh, N. (1986). Classroom intellectual composition and academic achievement. American Educational Research Journal, *23*, 357-374.
- Eder, D. (1981). Ability grouping as a self-fulfilling prophecy: A microanalysis of teacher-student interaction. Sociology of Education, *54*, 151-161.
- Finley, M. K. (1984). Teachers and tracking in a comprehensive high school. Sociology of Education, *57*, 233-243.
- Gamoran, A. (1987). The stratification of high school learning opportunities. Sociology of Education, *60*, 135-155.
- Gamoran, A. (1989). Measuring curriculum differentiation. American Journal of Education, *97*, 129-143.

- Gamoran, A. (In press). Access to excellence: Assignment to honors English classes in the transition from middle to high school. Educational Evaluation and Policy Analysis.
- Gamoran, A., and Berends, M. (1987). The effects of stratification in secondary schools: Synthesis of survey and ethnographic research. Review of Educational Research, 57, 415-435.
- Gamoran, A., Berends, M., and Nystrand, M. (1990). Classroom instruction and the effects of ability grouping: A structural model. Paper presented at the annual meeting of the American Educational Research Association, Boston.
- Gamoran, A., and Dreeben, R. (1986). Coupling and control in educational organizations. Administrative Science Quarterly, 31, 612-632.
- Gamoran, A., and Mare, R. D. 1989. Secondary school tracking and educational inequality: Reinforcement, compensation, or neutrality? American Journal of Sociology, 94, 1146-1183.
- Gamoran, A., and Nystrand, M. (1990). Tracking, instruction, and achievement. Paper presented at the World Congress of Sociology, Madrid.
- Goodlad, J. I. (1984). A place called school. New York: McGraw-Hill.
- Jackson, P. W. (1968). Life in classrooms. Chicago, Ill.: University of Chicago Press.
- Langer, J., and Applebee, A. (1987). How writing shapes thinking. Urbana: National Council of Teachers of English.
- Lucas, S. R. 1990. Course-based indicators of curricular track locations. Unpublished M.S. thesis, University of Wisconsin-Madison.
- Lucas, S. R., and Gamoran, A. (1991). Race and track assignment: A reconsideration with course-based indicators of track locations. Paper presented at the annual meeting of the American Sociological Association, Cincinnati.
- Mehan, H. (1979). Learning lessons. Cambridge, Mass.: Harvard University Press.

- Metz, M. H. (1978). Classrooms and corridors: The crisis of authority in desegregated secondary schools. Berkeley, Calif.: University of California Press.
- Meyer, J. W. (1980). Levels of the educational system and schooling effects. Pp. 15-63 in C. E. Bidwell and D. M. Windham (eds.), The analysis of educational productivity. Volume 2: Issues in macroanalysis. Cambridge, Mass.: Ballinger.
- Meyer, J. W., and Rowan, B. (1978). The structure of educational organizations. Pp. 78-109 in M. W. Meyer (ed.), Environments and organizations. San Francisco: Jossey-Bass.
- Murphy, J., and Hallinger, P. (1989). Equity as access to learning: Curricular and instructional treatment differences. Journal of Curriculum Studies, 19, 341-360.
- Moore, D. R., and Davenport, S. (1988). The new improved sorting machine. Madison, Wis.: National Center on Effective Secondary Schools.
- Newmann, F. M. (ed.). (In press). Student engagement and achievement in American secondary schools. New York: Teacher College Press.
- Newmann, F. M., Onosko, J., and Stevenson, R. (1988). Higher order thinking in high school social studies: An analysis of classrooms, teachers, students, and leadership. Madison, Wis.: National Center on Effective Secondary Schools.
- Nystrand, M. (1977). Language as discovery and exploration: Heuristic and explicative uses of language. Pp. 95-104 in M. Nystrand (ed.), Language as a way of knowing: A book of readings. Symposium Series/8. Toronto: The Ontario Institute for Studies in Education.
- Nystrand, M. (1992). Dialogic instruction and conceptual change. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Nystrand, M., and Gamoran, A. (1988). A study of instruction as discourse. Paper presented at the annual meetings of the American Educational Research Association.

- Nystrand, M., and Gamoran, A. (1991a). Student engagement, instructional discourse, and literature achievement. Research in the teaching of English, 25, 261-290.
- Nystrand, M., and Gamoran, A. (1991b). Student engagement: When recitation becomes conversation. Pp. 257-276 in H. Waxman and H. Walberg (eds.), Contemporary research on teaching. Berkeley, Calif.: McCutchan.
- Oakes, J. (1985). Keeping track: How schools structure inequality. New Haven, Conn.: Yale University Press.
- Oakes, J. (1991). Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science. Santa Monica: RAND.
- Oakes, J., Gamoran, A., and Page, R. N. (1992). Curriculum differentiation: Opportunities, outcomes, and meanings. Pp. 570-608 in P. W. Jackson (ed.), Handbook of Research on Curriculum. Washington, D.C.: American Educational Research Association.
- Page, R. N. (1991). Lower track classrooms: A curricular and cultural perspective. New York: Teachers College Press.
- Parsons, T. (1960). Structure and process in modern society. Glencoe, Ill.: Free Press.
- Perrow, C. (1986). Complex organizations: A critical essay. Third edition. New York: McGraw-Hill.
- Piaget, J. (1937). La construction du réel chez l'enfant. Neuchatel: Delachaux et Ciestlé.
- Rosenbaum, J. E. (1976). Making inequality: The hidden curriculum of high school tracking. New York: Wiley.
- Slavin, R. E. (1987). Ability grouping and achievement in elementary schools: A best-evidence synthesis. Review of Educational Research, 57, 293-336.
- Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. Review of Educational Research, 60, 471-499.

- Talbert, Joan. (1990). Teacher tracking: Exacerbating inequalities in the high school. Paper presented at the annual meeting of the American Educational Research Association, Boston.
- Thompson, J. D. (1967). Organizations in action. New York: McGraw-Hill.
- Vanfossen, B. E., Jones, J. D., and Spade, J. Z. (1987). Curriculum tracking and status maintenance. Sociology of Education, 60, 104-122.
- Walker, D. F., and Schaffarzick, J. (1974). Comparing curricula. Review of Educational Research, 44, 83-111.
- Wehlage, G. G., Rutter, R. A., Smith, G. A., Lesko, N., and Fernandez, R. R. (1989). Reducing the risk: Schools as communities of support. London: Falmer.
- Wehlage, G. G., and Smith, G. A. (In press). Building new programs for students at-risk. In F. M. Newmann (ed.), Student engagement and achievement in American secondary schools. New York: Teachers College Press.