Abstract Title Page
Not included in page count.

Title: Tracking and Student Achievement: The role of instruction as a mediator

Authors and Affiliations: Rebecca Anne Schmidt, Peabody School of Education at Vanderbilt University, Leadership Policy and Organizations program

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Most public schools and districts must face the problem of how to help low-achieving students and efficiently target resources, particularly in the face of accountability under No Child Left Behind. One policy that has been employed is grouping students into classrooms by their measured or perceived ability—a process known as tracking. Research has shown, however, that this practice disproportionately assigns minority and low-income students to low-track classes (e.g., Gamoran, 2009; Oakes, 2005) and may increase inequality between high and low-achieving students (e.g., Esposito, 1973; Gamoran, 1987; Oakes, 2005), without an average gain in student achievement in the school or district (e.g., Esposito, 1973; Gamoran, 2009; Kulik and Kulik, 1982; Slavin, 1990).

Despite the large number of reports examining outcomes in tracked and de-tracked classrooms, few large-scale studies have undertaken to look at the mechanism by which tracking may help or harm students, namely the quality of instruction. Opponents of tracking have argued that students who are placed in higher tracks have more qualified teachers and a more challenging classroom environment, which exacerbates existing achievement gaps; tracking proponents argue that separating students by ability allows teachers to more effectively target instruction to the diverse needs of students in their schools (Gamoran, 2009; Loveless 1999). In both cases instruction is the linchpin in making tracking or de-tracking work for students, so understanding the importance of teaching is paramount.

Yet, the vast majority of the studies in this area are small case studies using two or three teachers or schools (e.g., Boaler, 2006; Horn, 2006; McDermott, Rothenberg and Martin, 1995; Rubin, 2008; Watanabe, 2008). The few studies that have quantitatively examined mathematics instruction have generally relied on teacher self-report (Epstein and MacIver, 1992), focusing either on non-subject-specific teacher behaviors, such as classroom management, climate and teacher enthusiasm, or on curricular materials over instruction (Evertson, 1982; Oakes, 2005). To answer the question of whether instructional quality has an impact on the relationship between tracking and math achievement, we must start with a mathematics-specific definition of high quality instruction and carry that through a full quantitative mediation analysis.

The definition of high-quality mathematics instruction I will use begins with the National Council of Teachers of Mathematics (NCTM) standards, which are cited by many as what counts as "good" mathematics teaching and learning (e.g., Freeman and Crawford, 2008). These standards focus on the teacher knowing "what students know and need to learn," challenging all students and emphasizing conceptual understanding over procedural fluency alone (NCTM, 2000). The focus on conceptual understanding in particular links to prior research by Oakes (2005) and others qualitative researchers, who argued that a focus on explanation, justification and deeper understanding, in contrast to carrying out known procedures, prepares students for higher status jobs where more independent thinking is required.

Purpose / Objective / Research Question / Focus of Study:

This research focuses on the role of instructional quality in mathematics as a mediator between tracking and student achievement, and addresses the following research questions: 1) Are there measurable differences in instructional quality between teachers in tracked and untracked settings? 2) Between high- and low-track classrooms? 3) Do these differences mediate the relationship between track level (high- versus low-track) and student achievement?

Setting:

This analysis uses data from the Middle school mathematics in the Institutional Setting of Teaching (MIST) project at Vanderbilt University. MIST is an ongoing National Science Foundation-funded project examining the relationship between institutional supports, instructional practices and student learning in 30 middle schools in four large, urban districts. These districts were selected because they are undertaking instructional improvement initiatives in mathematics and have adopted inquiry-oriented curricula, such as the Connected Math Project (CMP). In each of the four districts, six to ten middle schools were selected in collaboration with central office staff to be representative of the district. Within these schools, teachers were randomly selected and recruited to participate in the study.

Population / Participants / Subjects:

As a part of the MIST project, nearly 120 teachers in four districts have been videotaped during instruction for two consecutive days, and their students' achievement data has been collected. This data included the current year mathematics achievement test results as well as two prior years' scores, which were standardized to state distributions by grade and year. Demographic data were also collected, including grade level, gender, race, free/reduced-price lunch, English Language Learner and Special Education status.

This sample of teachers is 66% female, with an average age of 39. The vast majority (91%) of teachers in MIST schools were fully certified, and 38% were non-white. The average years of experience in teaching mathematics was 9.4, but this varied greatly among the four districts. In district A, teachers had an average of 14 years of experience, while in district B, the average was less than eight. More than half of MIST teachers had a bachelor's degree as their highest degree, while 42% had a master's degree. This also varied by district: in district D nearly 70% of teachers hold a master's degree, while only 20% of district B teachers had a master's.

Intervention / Program / Practice:

"Tracking" is a word that has been used to describe a wide variety of policies and behaviors in schools. In this analysis it refers to sorting students at a classroom level by measured or perceived ability. Likewise, there are many ways to measure instructional quality. In this analysis, I will use a definition derived from the NCTM standards and measured by the Instructional Quality Assessment (IQA).

This study focuses on IQA's Academic rigor rubrics: Task Potential, Implementation, and Discussion. The Task Potential rubric is based on ratings of the cognitive demand of the task *as written*. High cognitive demand tasks are conceptualized as those with multiple solution paths and those that allow students to make connections between ideas and communicate their thinking. The Implementation rubric rates the level of cognitive demand *actually required* during the class period. The authors argue that teachers can and often do change the level of cognitive demand of a task over the course of the period. Finally, the Discussion rubric rates the "level of cognitive processes evident in the discussion" (Boston and Wolf, 2006, p. 13). High quality discussions require students to justify and compare their solution strategies to come to a deeper understanding of the underlying mathematics in the problem.

Research Design:

The first two research questions discussed above seek to establish the quantitative relationship between tracking and instructional quality. To address these questions, the dependent variables will be IQA ratings on the "Task Potential," "Implementation," and "Discussion" rubrics. Although the IQA scores are on an ordinal scale, I will dichotomize them to compare high quality (levels 3 and 4) to low quality (levels 1 and 2) instruction. The split between IQA level two and three is supported by the literature in the focus on the difference between instruction that emphasizes procedural learning without connections to the underlying mathematics and instruction that emphasizes conceptual learning (e.g., Horn, 2006; NCTM, 2000; Oakes, 2005; Stein et. al, 1996).

As mentioned above, the dataset used for this paper includes students of selected teachers and schools in four large, urban districts. This clustering of the data suggests a multi-level model approach to avoid the problem of correlated error terms (Raudenbush and Bryk, 2002). The models for these two research questions will require two-level models: teachers nested within schools. The tracking variables are teacher-level, as is the teacher's instructional quality.

For the third research question, I examine the role of instructional quality as the mechanism by which tracking affects achievement. This is illustrated in Figure 1. (Please insert Figure 1 here). While Baron and Kenny established the basic approach to testing for mediation in a linear regression approach, more recent research has extended this approach to working with multi-level data (Krull and MacKinnon, 1999, 2001). These researchers have pointed out that examining the impact of a group-level variable on an individual-level outcome will result in correlated error terms, as discussed above, violating a basic assumption of OLS. Therefore, this analysis will also use multi-level models with students nested within classrooms within districts.

Student achievement data will be entered at the first level, and track-level variables and teacher IQA ratings will be entered at the second (classroom) level. The district level will include only district-level random effects. Student and classroom control variables will also be included.

Data Collection and Analysis:

As discussed above, MIST participating teachers were observed in their classroom during to (ideally consecutive) days of instruction each year for four years. These classroom observations were scored using the Instructional Quality Assessment (IQA) each year. Additionally, the students' current and two prior years of achievement test results were collected from the district. As a part of the MIST study, data on tracking and track-level was collected through one-on-one interviews with teachers and principals, in which we asked whether classes were grouped by skill level (tracked) and what those levels were (track level). Principals provided an overall view of the courses offered at the school and how students were placed in them, while teachers provided the information on their particular classes. We then used this information to verify course files from the school and district. If a track level still could not be determined, we used the average prior achievement of the students in the course to assign a track level.

Findings / Results:

Analysis of the MIST data confirms the prior research finding that there is no statistically significant difference between tracked and untracked settings in achievement, but that there is a significant gap between high and regular/low-track students in the same schools. Controlling for prior achievement and student demographics, the size of this difference is nearly one standard

deviation when using OLS. Applying the multi-level model and controlling for classroom characteristics, the size of the difference is reduced to about 0.1 standard deviations, and there is significant dependency in the data: the intra-class correlation (ICC, or ρ) of the students within classes was 0.485, indicating that about 48.5% of the variation in student math scores is at the classroom level. Additionally, about 49% of the variation between classrooms is at the district level, with less than 1% at the school level. Therefore, the third level will be district, rather than school.

In response to the first research question, whether IQA scores vary by whether the grade level is tracked or not, I found that tracked classrooms had a significantly higher likelihood of implementing cognitively demanding tasks, but a significantly lower likelihood of cognitively demanding discussions. (Please insert Table 1 here)

In response to the second research question, whether IQA scores vary by track *level*, I found that high track classes have significantly *lower* likelihood of choosing rigorous tasks as written (potential of the task) than regular/low track classes. However, these classes had a higher likelihood of implementing rigorous tasks, and there was no significant difference between high and low-track classes in the likelihood of rigorous discussion. (Please insert Table 2 here)

In response to the third research question, whether IQA scores mediate the relationship between track level and student achievement, I found that less than 0.1% of the relationship between tracking and student achievement was mediated by Potential of the Task, and the indirect effect (the impact of tracking on student achievement through potential of the task) was not statistically significant. A slightly greater proportion of the relationship was mediated by Implementation of the Task (about 3%), but the indirect effect was still not statistically significant. Likewise, more than 10% of the relationship between tracking and achievement was mediated by discussion scores, but the indirect effect was not statistically significant ($p=0.16$). I also examined the mediation effect of these three rubrics combined. When combined, the three measures mediate approximately 12.5% of the relationship between track level and student achievement, but the indirect effect is still not statistically significant ($p=0.13$).

Conclusions:

This analysis found only a small mediation effect of instructional quality on the relationship between track level and student achievement. This indicates that other variables are driving this relationship besides a difference in rigorous tasks and mathematical discussions. Potential competing theories include teacher characteristics (which were not controlled here) and classroom characteristics such as behavior management and classroom climate, which are not measured by the IQA.

These analyses suffer from several barriers to inferring causality. First, this sample included only a small number of classrooms in a given year: about 120 in each year. Second, the same teachers were not observed in high- and low-track classes, so there is the potential for spurious variables in comparing across teachers. Finally, the definition of instructional quality stems from one put forth by NCTM, and may not be aligned with what is tested on the state tests.

Appendices

Not included in page count.

Appendix A. References

Baron, R.M. and Kenny, D.A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic and Statistical Considerations, *Journal of Personality and Social Psychology, 51*(6): 1173 – 1182.

Boaler, J. (2006). How a Detracked Mathematics Approach Promoted Respect, Responsibility, and High Achievement. *Theory into Practice, 45*(1): 40-46.

Boston, M. and Wolf, M.K. (2006). Assessing Academic Rigor in Mathematics Instruction: The Development of the Instructional Quality Assessment Toolkit, *CSE Technical Report 672, Center for the Study of Evaluation.*

Epstein, J.L. & MacIver, D.J. (1992). *Opportunities to learn: Effects on eighth graders of curriculum offerings and instructional approaches* (Report No. 34). Baltimore, MD: Center for Research on Effective Schooling for Disadvantaged Students.

Esposito, D. (1973). Homogeneous and Heterogeneous Ability Grouping: Principal Findings and Implications for Evaluating and Designing More Effective Educational Environments, *Review of Educational Research, 43*(2): 163 – 179.

Evertson, C.M. (1982). Differences in Instructional Activities in Higher- and Lower-Achieving Junior High English and Math Classes, *The Elementary School Journal, 82*(4): 329 – 350.

Freeman, B. and Crawford, L. (2008). Creating a Middle School Mathematics Curriculum for English-Language Learners, *Remedial and Special Education, 29*(1): 9 – 19.

Gamoran, A. (1987). The Stratification of High School Learning Opportunities, *Sociology of Education, 60*(3): 135-155.

Gamoran, A. (2009). Tracking and Inequality: New Directions for Research and Practice. *WCER Working Paper Number 2009-6* http://www.wcer.wisc.edu/

Horn, I.S. (2006). Lessons Learned from Detracked Mathematics Departments. *Theory into Practice, 45(1):* 72-81.

Krull, J.L. and MacKinnon, D.P. (1999). Multilevel Mediation Modeling in Group-Based Intervention Studies, *Evaluation Research, 23*(4): 418 – 444.

Krull, J.L. & MacKinnon, D.P. (2001). Multilevel Modeling of Individual and Group Level Mediated Effects, *Multivariate Behavioral Research, 36*(2): 249 – 277

Kulik, C.C. and Kulik, J.A. (1982). Effects of Ability Grouping on Secondary School Students: A Meta-Analysis of Evaluation Findings, *American Educational Research Journal, 19*(3): 415-428.

Loveless, T. (1999). *The tracking wars: State reform meets school policy.* Washington, DC:Brookings Institution Press

McDermott, P., Rothenberg, J. & Martin, G. (1995). "Should We Do It the Same Way?" Teaching in Tracked and Untracked High School Classes. *Paper Presented at the Northeastern Educational Research Association 26th Annual Conference, October 25-27, 1995*

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics.* Reston, VA: Author.

Oakes, J. (2005). *Keeping Track: How schools structure inequality, Second Edition*. New Haven, CT: Yale University Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods second edition.* London: Sage.

Rubin, B. (2008). Detracking in Context: How Local Constructions of Ability Complicate Equity-Geared Reform. *Teachers College Record, 110(3):* 646-699.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis, *Review of Educational Research, 60*(3): 471—499.

Stein, M.K., Grover, B.W. and Henningsen, M. (1996). Building Student Capacity for Mathematical Thinking and Reasoning: An Analysis of Mathematical Tasks Used in Reform Classrooms, *American Educational Research Journal, 33*(2): 455 - 488.

Watanabe, M. (2008). Tracking in the Era of High-Stakes State Accountability Reform: Case Studies of Classroom Instruction in North Carolina. *Teachers College Record, 110(3)*: 489-534.

Appendix B. Tables and Figures

Figure 1

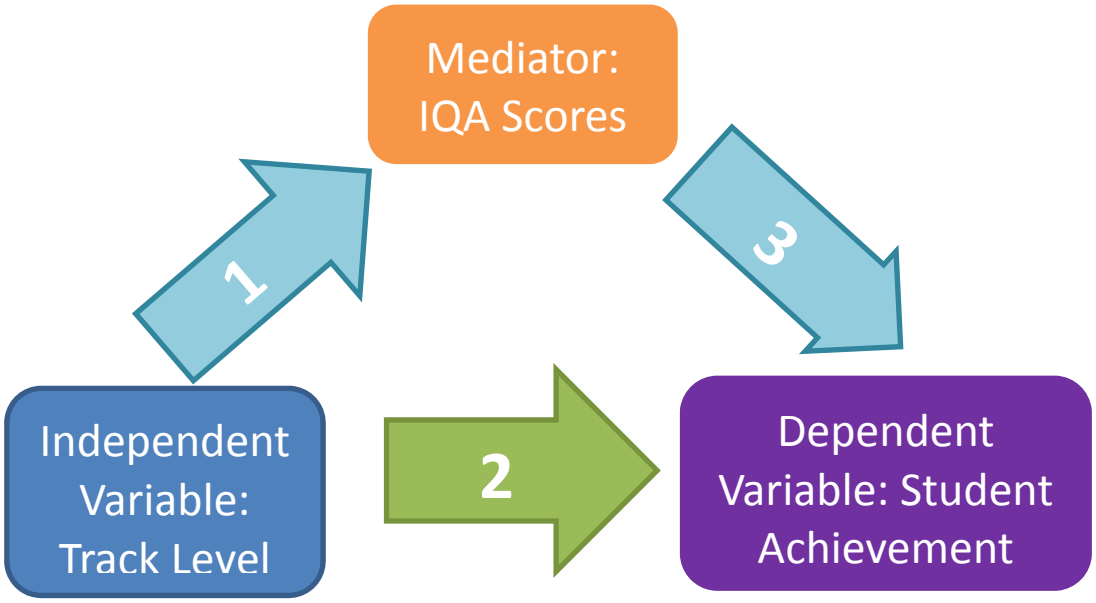

Table 1: Relationship between Tracking and IQA

	Potential		Implementation		Discussion	
	b	SE	b	SE	b	SE
Tracked	**-0.20**	**(0.11)**	**1.39*****	**(0.13)**	**-0.43*****	**(0.13)**
Year 2	-0.75***	(0.07)	-0.62***	(0.08)	-0.36***	(0.09)
Year 3	-0.59***	(0.07)	-0.06	(0.07)	-0.19*	(0.09)
District B	-0.98	(0.75)	-1.26	(1.04)	0.53	(1.21)
District C	-2.33**	(0.78)	-2.66*	(1.11)	-0.87	(1.27)
District D	-0.70	(0.75)	-1.74	(1.04)	1.82	(1.21)
Class is 7th grade	-0.00	(0.09)	-0.93***	(0.09)	-1.40***	(0.12)
Class is 8th grade	0.22**	(0.08)	-0.40***	(0.08)	-0.40***	(0.09)
Pct FRL	0.12	(0.23)	-2.57***	(0.25)	-3.10***	(0.27)
Pct LEP	1.93***	(0.37)	1.99***	(0.37)	-0.28	(0.45)
Pct SPED	0.33	(0.27)	2.20***	(0.28)	-3.97***	(0.41)
# of test-takers in class	-0.02***	(0.00)	-0.05***	(0.01)	-0.04***	(0.01)
Pct Minority	0.32	(0.34)	-2.12***	(0.37)	3.57***	(0.46)
Constant	2.63***	(0.57)	4.54***	(0.76)	-0.87	(0.90)
Level 2						
Constant	0.36*	(0.16)	0.72***	(0.16)	0.85***	(0.17)
Observations	7309		7309		7309	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Relationship between Track Level and IQA

	Potential		Implementation		Discussion	
	b	SE	b	SE	b	SE
High Track	**-0.31****	**(0.11)**	**0.57*****	**(0.13)**	**1.23*****	**(0.18)**
Year 2	-0.68***	(0.09)	-0.87***	(0.10)	-0.21	(0.12)
Year 3	-0.75***	(0.09)	-0.26**	(0.09)	0.31*	(0.12)
District B	-1.13	(1.09)	-1.87	(1.71)	2.49	(1.95)
District C	-2.66*	(1.14)	-4.54*	(1.87)	-0.72	(2.08)
District D	-1.05	(1.27)	-2.16	(2.00)	4.98*	(2.26)
Class is 7th grade	0.68***	(0.11)	-1.19***	(0.12)	-2.91***	(0.20)
Class is 8th grade	0.72***	(0.10)	-0.32**	(0.10)	-0.47***	(0.12)
Pct FRL	-0.73*	(0.35)	-3.55***	(0.43)	-5.48***	(0.54)
Pct LEP	3.59***	(0.44)	3.35***	(0.43)	-1.66**	(0.59)
Pct SPED	-0.53	(0.31)	2.54***	(0.33)	-4.24***	(0.49)
# of test-takers in class	-0.01	(0.01)	-0.06***	(0.01)	-0.09***	(0.01)
Pct Minority	0.09	(0.46)	-0.91	(0.49)	5.01***	(0.64)
Constant	2.83**	(0.88)	6.30***	(1.30)	-1.60	(1.54)
Level 2						
Constant	0.69***	(0.20)	1.15***	(0.21)	1.24***	(0.20)
Observations	5168		5168		5168	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$