國中能力分班對學習成就的影響：

結合傾向分數與多層次的分析

關秉寅
國立政治大學社會學系
soci1005@nccu.edu.tw

本論文發表在 2012 年 11 月 25 日於東海大學社會學系及台灣社會學會主辦之「2012 台灣社會學會年會暨國科會專題研究成果發表會」。本論文研究為行政院國家科學委員會專題研究計畫（NSC 99-2410-H-004-166）的部分成果。

The Effects of Tracking on Academic Achievement:

Evidence from Junior High Students in Taiwan

Ping-Yin Kuan
Associate Professor
Department of Sociology
National Chengchi University
E-mail: soci1005@nccu.edu.tw

A paper prepared for the Annual Meeting of Taiwan Sociological Association, Nov. 24-25, 2012, Tunghai University, Taichung City, Taiwan. The present study is supported by the funding of the National Science Council, Taiwan, R. O. C. (NSC 99-2410-H-004-166).

摘要

本研究的目的是結合傾向分數配對（propensity score matching）方法與多層次模型（mutlti-level modeling）的分析方法，分析具台灣地區代表性之「台灣教育長期追蹤資料庫」（以下簡稱TEPS）追蹤樣本現場使用版，以瞭解國中時期選入前段班或好班之學生生在國中時期學習成就上的差異。

能力分班對於中小學時期學習成就的效果在國外研究甚多，但並無定論。沒有定論的原因之一是因為傳統分析方法無法有效控制未觀察到之因素對能力分班效果的影響。在台灣，利用具全國代表性樣本在此議題上的分析則仍付之闕如。具體而言，本研究企圖回答的問題是：國中時期進入前段班或好班的學生，其學習成就是否會有如一般所言的正面效果？

本計畫針對TEPS資料選出國一時並無能力分班，但國二或國三時進入或未進入前段班者，以各種分析策略評估國二及國三均在前段班、只有國二在前段班，以及只有國三在前段班者之學習成就是否與未進入者不同。分析的策略包括使用一般線性迴歸模型、固定效果模型、隨機效果模型，以及結合這些模型及傾向分數配對之多層次傾向配對模型的分析。

各類分析策略分析的結果顯示，進入前段班的效果會隨分析模型之不同可有頗大的差異。但整體趨勢言，國二及國三進入前段班，以及只有國三進入前段班者對其學習成就有正面影響，但只有國二在前段班者，則對學習成就的影響可能不顯著，或相對於其他兩類進入好班者言，其正面效果較小。

關鍵詞：能力分班、台灣教育長期追蹤資料庫、傾向分數配對、多層次模型、固定效果模型、隨機效果模型

Abstract

The purpose of the present research is to assess the causal effects of tracking on junior high students' academic achievement in Taiwan. In this study, tracking refers to a school's practice of separating students into different classes, courses, or course sequences. In order to overcome the possible selection bias at either student or school level, the present research uses different methods of estimation to evaluate the causal effect of tracking in junior high. These methods include fixed effects and random effects models as well as combination of these methods with the method of propensity score matching.

The research focused on students who were not tracked in the 7th grade and then assigned to high-ability groups in either 8th or 9th grade or in both grades. Hence the research question is: Would junior high students who assign to high-ability groups benefit from the tracking in terms of 9th grade academic achievement?

This study uses data gathered by Taiwan Education Panel Survey (TEPS) in 2001 and 2003. The analytical sample of the present study is those who were not tracked in the 7th grade (N=12,513). This sample selection excludes those who claimed to be tracked since 7th grade. By doing this, the present study can be certain about the causal order between the tracking practices occurred after 7th grade and the observed outcomes.

The main findings of this study are: (1) In Taiwan, junior high students' own motivation, effort, and ability seem to be major factors contribute to their selection into high-ability groups; (2) In general, students being assigned to high-ability groups for two years tend to gain slightly more than those being assigned to the same kind of groups at 9th grade only. Those who assigned to the high-ability groups at 8th grade and then exited the high ability groups would have the least or even no gain in their academic performance in 9th grade, and (3) Different models of estimation offer quite different estimates of the causal effect of being assigned to high-ability groups on academic achievement.

Keywords: tracking, selection bias, fixed effects, random effects, propensity score matching, Taiwan Education Panel Survey

The Effects of Tracking on Academic Achievement:
Evidence from Junior High Students in Taiwan

Ping-Yin Kuan

1. Introduction

Tracking in the form of separating students into different classes, courses, course sequences, and schools based on their achievement, objectively or subjectively evaluated, is a common education practice in many industrialized countries (Hanushek & Woßmann, 2006; Huang, 2009; Schofield, 2010). Taiwan is of no exception. Despite the practice of tracking in the form of grouping students into separate classes at the stage of junior high (7th to 9th grade) is prohibited by the Ministry of Education in Taiwan and many advocates for educational reform protest strongly against the practice, junior high schools still use many forms of tracking, both legally and illegally, to enhance student's competitiveness in the high-stake senior-high school entrance examination (Lin, 2010). The 9th graders of the junior high in Taiwan would need good test grades to enter elite senior high schools. Schools operate the practice of tracking under various names, like gifted classes, to circumvent the restriction of law. Since the educational system allows tracking between schools at the stage of senior high, the present research focuses on the impact of the tracking practice within school at the stage of junior high.

Tracking has been a controversial and much investigated issue in the U. S. (e.g., Ansalone, 2010; Loveless, 1999; Oaks, 1985). Quite a few empirical studies outside of the U. S. or cross-countries comparisons on the impact of tracking or ability grouping on student performance can also be found (see Schofield, 2010; Van de Werfhorst & Mijs, 2010, for reviews). The key issue of tracking is whether this widely adopted educational practice would have impact on students' achievement, which in turn may lead to widened inequality among students tracked into different ability groups or types of schools. To contribute to the growing international literature on the consequences of tracking, the purpose of this research is to focus on the within school tracking practice at the stage of junior high in Taiwan, which has not been much studied so far, and to explore the impact of the tracking practice on students' academic achievement in their final year at junior high.

Does within school tracking contribute to students' academic achievement? Depending on the research design and the quality of the data, the effect of tracking may be positive, negative, or neutral (Kulik, 2004). Moreover, the effect of tracking differs from country to country (Gamoran, 2009). Tracking has been found to have a positive impact on student performance in countries with high-stake testing, such as South Korea and Israel. Taiwan is also a country with high-stake testing. The educational policy in Taiwan also offers a standard curriculum guideline which dictates the content of teaching and learning in junior high. Hence, the effects of tracking in Taiwan may not be as negative as the Ministry of Education or educational reformers in Taiwan have contented.

The following sections will first review briefly the literature about the possible effects of within school tracking on students' academic performance, as well as the conditions that may bring about the effect of tracking. Following the literature review is the section that describes the research design including data, methods, and measures used to assess the effect of tracking in this study. Finally, findings will be presented and discussed.

2. Tracking and its causal effects on students’ academic performance

In the present study, tracking refers to a school’s practice of separating students into different classes, courses, or course sequences based on their achievement. Previous research findings about the effects of tracking have been inconsistent. The literature has debated about the impact of tracking and the proper method should be used to evaluate the effect of tracking. Both qualitative and quantitative methods have been used to evaluate the effect of tracking. Studies with different designs and methods have shown tracking to have either positive, negative, or non-significant effect on students’ academic performance.

2-1 Does tracking work?

Advocates of tracking believe that tracking benefits students’ academic performance and detractors, on the other hand, believe that tracking would increase the achievement gap between tracked students.

Research on tracking in U.S. and U.K. has consistently shown that tracking was positive to the students who were assigned to high ability tracks and was negative to the students assigned to low ability tracks (see Gamoran, 2010 for review). Hence, tracking increases the achievement gap between high tracks and low tracks, and enlarges inequality. A few meta-reviews, however, found that within school tracking or ability grouping in general had no significant impact of students' achievement (Kulik, 2004; Slavin, 1990). Hallinan (1994), on the other hand, has argued that tracking would decrease inequality if the criterion of tracking was students’ performance or grades.

Studies outside of the U. S. have shown that tracking could lessen achievement inequalities. For instance, Kim, Lee & Lee (2008), after comparing the academic performance of high school students enrolled in either the mixing and sorting system in South Korea, found that sorting system helps students increase their test scores.

Broaded (1997) investigated Taiwan’s situation and found that students benefited from tracking and the negative effect of disadvantage family has been lessened.

Gamoran (2010) maintained that the key condition that brings about the positive effect of tracking in Taiwan or South Korea is the high-stakes exam. This kind of exams leads governments to standardize the curriculum and assessment. Facing the exams, both teachers and students are committed to gaining high scores. Consequently, to be successful in the exam is a strong incentive to both students and teachers, and the incentive results in favorable influence of the tracking.

Reviewers of empirical studies of tracking or ability grouping also pointed out that one possible mechanism for tracking to be positive to students' performance or to increase achievement gap between tracked groups is for ability grouping to be accompanied with curriculum differentiation (Kulik & Kulik, 1992; Kulik, 2004; Schofield, 2010; Slavin, 1990). Tracking or ability grouping without providing students with different levels of difficulty or providing them with different programs or instruction would not produce significant impact on students' performance.

2-2 The limitations of the past research

Kulik (2004) in his extensive review of prior research on tracking has divided the research into three kinds in terms of the research design: experimental, correlational, and ethnographic. Different research designs may lead to divergent findings of the effect of tracking mentioned earlier. Kulik (2004) believed that the most reliable evidence comes from experimental research, because of its elegant design and method. Since correlational analyses can’t control all the factors which may affect both student achievement and track placement, their causal effect is inconclusive. Ethnographic studies usually focus on their cases, general limited number, and their results could not be generalized easily to other cases.

As Gamoran (2010) has argued, to investigate the effects of tracking and its causal relationship, we must distinguish the effects of track assignment from the effects of preexisting differences among students and its effects after tracking. Quite often students are tracked to different classes or groups by factors other than achievement. These factors include students' motivation, aspiration, gender, race or ethnicity, and socioeconomic status. Teacher and parental expectation and judgment also matter (Gamoran & Berends, 1987). If we can’t discern these two causes, we could make commit the mistake of selection bias, which could mislead our findings. Most of prior quantitative studies of tracking, however, used cross-sectional data which may not give sufficient information to account for selection bias fully. Longitudinal data, on the other hand, because of repeated observations of the same individuals, can be used more effectively to control selection bias and catch individual variation.

The investigation of the effect of tracking is further complicated by the fact the practice of tracking is offered by the school. Hence, to assess the effect of tracking, we not only have to control for individual preexisting differences, but also school differences. Schools with limited resources and students of broad range of academic abilities within the class and the school may prompt teachers and principals to favor within school tracking and believe that tracking may make classroom life more manageable (Ansalone & Biafora, 2004; Biafora & Ansalone, 2012).While methods with hierarchical linear modeling can be used to include the school-level variables in our examination of the impact of tracking (e.g., Gamoran, 1992), Arpino & Mealli (2008) points out that there is also the problem of selection bias at the cluster or group level, if one or more cluster level variables affect the selection of individuals into treatment groups.

With the awareness of the limitations of previous research in mind, the present study uses multi-level longitudinal data as well as the method of fixed effects model, random effects model, and propensity score matching (PSM) as well as the combination of PSM and fixed effects model to deal with the possible problem of preexisting differences at both student and school level.

3. Research Design

Previous inconsistent findings about the effects of tracking may be due in part to the use of observational data without properly controlling for selection bias or endogeneity. Regarding the assessment of the causal effect of tracking, the selection bias or endogeneity may occur at either the individual student level or the school level. In order to overcome the possible selection bias at both levels, we explore the advantage of a longitudinal data set and different estimation models to assess the causal effect of tracking in junior high.

3-1 Sample

The present study uses data gathered by Taiwan Education Panel Survey (TEPS) in 2001 and 2003. TEPS sampled 333 junior high schools. Within each school, TEPS in principle sampled 4 classes and within each class, it randomly selected 15 students (Chang, 2009). Between 2001 and 2003, the number of classes surveyed increased from 1,244 to 1,938. The increase to large extent is related to the practice of tracking or ability grouping. In 2001, TEPS surveyed 20,005 7^{th} grade students. In 2003 follow-up survey, the number of students surveyed was 19,088.

This study used the on-site version of the TEPS data[1] which offers more detailed

[1] The authorization code for using the on-site version of the TEPS data is TEPS2A002097.

information about students' teachers, classes, and schools. The access of the on-site version has to be approved by the Survey Research Center, Academia Sinica (Chang, 2009). The analytical sample of the present study is those who were not tracked in the 7th grade (N=12,513). This sample selection excludes those who claimed to be tracked since 7th grade. By doing this, the present study can make sure that the causal order between the tracking practices occurred after 7th grade and the observed outcomes is proper. The sample also excludes those who studied private schools or changed schools between two waves. The sample size of students who were assigned to high-ability group in both 8th and 9th grade is 547, those in 8th grade high-ability group only is 587, and those in 9th grade high-ability group only is 1,863. These students were compared to students who had no tracking experiences or had attended schools with no tracking system (N = 9,516).

3-2 Methods

For the present study, several different strategies are employed to control for possible selection bias or endogeneity at either the student level or the school level. These models used include fixed effects and random effects models with either students or schools as clusters and the method of propensity score matching (PSM). The main difference between the fixed effects and random effects model is the assumption about the relationship between unobserved variables and explanatory variables included in the model. For random effects model, the unobserved variables are assumed to be a set of random variables and to be independent of explanatory variables included in the model. The fixed effects model, on the other hand, treating the unobserved as a set of fixed numbers while allows the unobserved variables to be correlated with explanatory variables included in the model (Allison, 2009).[2] If there are unobserved variables at both student and school level and the relationship between the unobserved at either level and explanatory variables included in the model is not independent, then using either fixed or random effects model with either students or schools as clusters may not be able to handle the problem of endogeneity effectively (Ebbes, Böckenholt, & Wedel, 2004). A possible alternative then is to combine either fixed or random effects model with PSM (Arpino & Minealli, 2011; Buscha et al., 2011).

When using non-experimental or observational data to evaluate causal effects, PSM developed under the framework of counterfactual causal inference, is gaining popularity to overcome the problem of nonrandom assignment (Guo & Fraser, 2010). Other than using PSM to match covariates at both individual level and school level directly, this study also combines PSM with fixed effects models. The latter PSM

[2] This study used xtreg, a Stata command, to perform fixed effects and random effects regression.

strategy can have two approaches. The first approach is using fixed effects models to obtain propensity scores of being tracked first, and then uses one-to-one and one-to-five matching to estimate the treatment effect of being tracked at either 8th, 9th, or both grades.[3] The fixed effects model used to estimate PS of being tracked is essentially the same as estimating the probability of being tracked with schools as dummy variables along with other student level covariates. Arpino and Mealli's (2011) simulation demonstrated that comparing with models ignored the cluster-level covariates, multi-level PSM models that used random or fixed effects models for the estimation of propensity score performed fairly well in dealing with the omitting cluster-level covariates. In particular, the PSM using the fixed effects models has the best performance under different scenarios of unobserved cluster-level variables.

The second approach of combining PSM and fixed effects model is to use the PSM method to match students being tracked and students who had no tracking experience during junior high, and then uses fixed effects regression to estimate the treatment effect of being tracked at either 8th, 9th, or both grades.[4] Matching students being tracked with those who were not tracked allow us to have sufficiently comparable groups for further analysis. Combining PSM and fixed effects regression enable us to relax the linearity assumption, relative to standard fixed effects regression, in estimating differences in outcomes over time (Buscha, et al., 2012).

The method of propensity score matching per se can be used to estimate three different types of causal effect: the average treatment effect on the treated (ATT), the average treatment effect on the untreated (ATU), and the average treatment effect of both treated and untreated (ATE) (Morgan & Winship, 2007). The treated in the present study would be students who were assigned to high-ability classes. The present study focuses on estimating the average treatment effect on the treated (ATT) by comparing separately the academic achievement of 9th of students who were assigned to high-ability track in both 8th and 9th grade, in 8th grade only (N = 587), or in 9th grade only. Since TEPS data offer no specific information about students who were assigned to middle/low-ability tracks since the 8th grade, no comparable comparison can be made between students of middle/low ability tracks and students who were not tracked or who were assigned to high-ability tracks.

For the purpose of understanding the possible bias caused by omitting important variables at individual or school level, the baseline model of the current study is an OLS regression model, which includes only individual-level covariates without the 7th grade ability score. The OLS regression is usually faulted by the possibility of not

[3] This study used xtlogit to perform fixed effects logistic regression and psmatch2 to perform PSM analysis.
[4] This study used xtreg to perform fixed effects regression or random effects regression.

including important control variables in the model, which in turn may biases the estimation of the causal effect of the focal explanatory variable. To explore the impact of omitted important control variables under the framework of OLS regression, an OLS model that includes the 7th grade ability score as well as an OLS model that further includes class-level and school-level means and standard deviations of academic achievement are specified as well.

The status of being tracked in either 8th or 9th grade is based on students' self-reports. Since TEPS does not offer sufficient information to identify schools' tracking practices objectively, students' reports is the only way to learn their tracking status in junior high. Lucas and Gamoran (2002) suggested that student-reported tracking status might capture the social-psychological dimension of tracking and be a useful indicator of students' academic attitudes and behaviors (see also Gamoran, 1987). Since tracking as defined by the present study is a school’s practice, it is reasonable to assume that most of matching done would mainly be matching across schools which either implement tracking or not.

3-3 Measures

The outcome variable of the present study is a transformation of the 9th-grade IRT ability score, which is an indicator of student's academic achievement. The variable is derived from the TEPS test results using 3-P Item Response Model. For the ease of interpretation, this outcome variable as well as the 7th-grade achievement scores is transformed into NCE-like scores. The 7th-grade achievement score has a mean of 50 and a standard deviation of 10. The transformation of the 9th-grade achievement scores uses the 7th-grade scores as the basis and has a mean of 56.923 and a standard deviation of 12.166. In other words, in comparison with their achievement 2 years ago, the 9th graders on average gained about 7 points and the dispersion of their performance also increases somewhat. The range of the 9th grade achievement scores is from 24.742 to 94.198.

Other than the 9th grade ability score, three dummy-coded explanatory variables indicating the status of being tracked in 8th grade, 9th grade, or both 8th and 9th grade are obtained from the student survey data gathered by TEPS in 2003. All other control or matching variables included in the various models are obtained from the student survey data and the parent survey data gathered by TEPS in 2001 when students were 7th graders. These control or matching variables are of three kinds: individual-level variables, class-level variables, and school-level variables.

The individual-level control/matching variables consist of student’s gender, ethnicity, parental levels of education, parental occupations, family monthly income, family structure, student’s own educational expectation, parents’ educational expectation, hours of attending cram schools at the 7th grade, and student's ability score at 7^{th} grade.

The following is a more detailed description of the construction of these variables:

(1) Male: student's gender is coded as a dummy variable with male as 1 and female as 0.

(2) Ethnicity: Five ethnic groups are constructed according to parents' answer about their ethnicity. They are Minnan, Hakka, Mainlander, Aborigine, and other ethnicity. The ethnic groups identified in this paper are conventional social differentiation in Taiwan. Minnan and Hakka are groups of Han Chinese residing in Taiwan for more than three generations and are differentiated linguistically. Mainlanders are descendants of Chinese who migrated to Taiwan around 1950s. Indigenous peoples have different cultures and languages from those of the Chinese. Minnan is the dominant linguistic-ethnic group in Taiwan.

(3) Parental levels of education: Six categories of education that indicate mainly the highest educational level that parents had earned are constructed. The six categories are junior high or below, senior high, junior college, 4-year college, graduate school, and other.

(4) Parental occupations: Six categories of occupation which indicate mainly the highest skill level of parents' occupations are constructed. The six categories are farmer or no skilled worker, semi-skilled or service worker, clerical worker, semi-professional, professional, and other.

(5) Family monthly income: The monthly family income is divided into under NT$20,000, NT$20,000 to NT$49,000, NT$50,000 to NT$99,999, NT$100,000 to NT$149,999, NT$150,000 to NT$199,999, and NT$200,000 or above.

(6) Family structure: Four types of family structure are constructed, which are living with both biological parents, single-father headed family, single-mother headed family, and all other types.

(7) Student's own educational expectation: This variable is coded into five categories of educational expectation, which are expectation of getting a high school diploma or below, getting a junior college degree, getting a 4-year college degree, getting a graduate degree, and other.

(8) Parental educational expectation: Same as the previous variable, five categories of parental educational expectation are constructed.

(9) Hours of attending cram schools per week at the 7th grade: Hours of attending cram schools per week at the 7th grade is considered as an indicator of student's academic effort. The variables is coded into five dummy variables which are none, less than 4 hours, 4 to less than 8 hours, 8 hours to less than 12 hours, and 12

hours or more.

(10) Student’s ability score at 7th grade: Since TEPS first surveyed the 7th graders in the middle of the first semester in 2001, we believe students’ ability score measured at that time in large part could be attributed to their ability gained before entering junior high. Every student's 7th grade ability score is centered by the mean ability score of the class attended by the student.

The class-level variables are the means and the standard deviations of the ability scores at the class level. The mean scores at the class level are centered by the school means. The school-level variables are the means and the standard deviations of the ability scores at the school level. These class-level and school-level variables can be considered as outcomes of various characteristics of schools and classes, such as school size and resource, learning climate, teaching quality, diversity and average socioeconomic status of the school district, etc.

Whether schools are identified by the Ministry of Education as the schools in remote areas, which tend to be small in size and students need to travel far away from home to attend schools, is also included in the study as a dummy variable. These schools also have less resource and fewer teachers, who may need to cover several different subjects. Table 1 offers summary statistics of all the variables used for the present study.

4. Findings

4-1 The propensity of being tracked to high-ability groups

Based on three separate binary logistic regressions, Table 2 presents the propensity of being tracked to high-ability groups in both 8th and 9th grade, in the 8th grade only, and in the 9th grade only. Based on the odds ratios of predictors in three regression models, variables related to student's own educational expectation, hours of attending cram schools per week, the 7th grade ability score, and the mean ability score at school level have fairly consistent impacts on student's chance of being tracked to high-ability classes in either 8th or 9th grade. Specifically, students who expected to have some college education, who spent some time to attend cram schools in the 7th grade, and who have higher 7th grade ability scores are having better chances of being assigned to the high-ability groups in the 9th grade, the last year in the junior high and the year to prepare for the senior high entrance examination. Table 2 also shows that students who expected to attend junior college, spent between 4 to 12 hours in cram schools, and had high 7th grade ability scores would better their chances of being assigned to high-ability groups in either 8th or 9th, or both grades. In short, it seems that variables reflecting student's academic motivation, effort, and ability in the 7th grade are important and consistent factors for students to be selected into high-ability groups.

[Table 2 is about here]

Variables related to student's gender and family backgrounds do not show any consistent pattern in predicting their chances of being assigned to high-ability groups. In the case of being assigned to high-ability groups in 8th grade only, the only background factor that decreases the chance is being Mainlanders. A few more background factors show impacts on chances of being assigned to high-ability groups in other two situations. In the case of being assigned to high-ability groups in both 8th and 9th grade, being male, having parents whose highest educational level is graduate school, having monthly family income of NT$150,000 to less than NT$200,000, and living with single-father all would lessen the chance. Interestingly, students who were economically deprived (family having monthly income less than NT$20,000) had a better chance of being in high-ability classes for both 8th and 9th grade. The background factors that depress student's chance of being assigned to high-ability groups in the 9th grade only are having parents whose highest level of education were junior college, family monthly income of NT$100,000 to NT$149,999 or over NT$200,000. Being Hakka, however, would increase the chance of entering high-ability groups in the 9th grade only.

At the class and school level, it is interesting to see that schools having higher mean ability scores tend to be less likely to adopt the tracking practice. Moreover, students assigned to high-ability groups in both 8th and 9th tended to come from classes or schools having greater diversity in mean ability scores. It seems that schools performed less well on average or having greater diversity in students' academic performance believed that tracking could enhance students' learning and academic competitiveness. Table 2 also shows that even facing the challenge of senior high entrance examination, remote schools in Taiwan were less likely to tracked their students in the 9th grade.

In summary, junior high students' own motivation, effort, and ability seem to be major factors contribute to their selection into high-ability classes in Taiwan. Whether or not a school would track its students is also related to the school understanding about its students' performance on average and the diversity of the performance at both class and school level. Finally, contrary to findings in some industrialized countries, such as the U. S., it seems that students came from privileged background in Taiwan do not gain any advantage in getting into high-ability classes. Instead, a few pieces of evidence suggest that students who were economically or culturally disadvantaged might have more chances in being assigned to high-ability classes in junior high.

4-2 The causal effects of being tracked to high-ability groups

Do junior high students assigned to high-ability groups in either 8th or 9th grade benefit from the tracking practice? The results shown by Table 3 indicate that the answer to the question depends on the grade that students were assigned to high-abilitygroups. Table 3 shows that in general, students being assigned to high-ability groups for two years (both 8th and 9th grade) tend to gain slightly more than those being assigned to the same kind of groups at 9th grade only. Those who entered into the high-ability classes at 8th grade and then exited would gain the least or have no gain, and may even have a negative impact on their academic performance. Table 3 also shows fairly clearly that depending on models, the estimated size of the gain or loss differs considerably.

[Table 3 is about here]

The baseline model to be compared by all other models shown in Table 3 is an OLS regression model (OLS1) that includes only student level covariates and without student's 7th grade ability score, which is an indicator of prior ability and as mentioned earlier an important factor influencing student's chances of being assigned to high-ability classes. Without student's prior ability and variables of class and school level included in the model, OLS1 estimation of causal effects of tracking is highly possible to be biased. Similarly, causal effects estimated by random effects or fixed effects models that attempt to control the unobserved only at one level, either student or school level, may also be biased. RE1, RE2, FE1, and FE2 are such models. The results shown in Table 3 indicate that in comparison with other OLS models that include student's prior ability and covariates at class and school level and PSM models that attempt to control for unobserved at both student and school level, effects of tracking estimated by OLS1, RE1, RE2, FE1, and FE2 are in general much bigger. For instance, RE1 which took into account the unobservable at student level offered the largest effect of being assigned to high ability groups at both 8th and 9th grade. The effect is 10.543 which is close to .9 standard deviation of the 9th grade ability score of the whole sample (sd = 12.166). The estimate of OLS1 of the same tracking status is 6.096, which is the smallest among these possibly biased models. This estimate is also about one half of the standard deviation of the 9th grade ability score. All these five possibly biased models also indicate positive effect of being assigned to high ability groups in 9th grade only. The positive effect, however, is somewhat smaller than that of being assigned to high ability groups in both 8th and 9th grade.

FE1, in this case, offers the largest effect, which is 9.671. As to the effect of being assigned to high ability groups at 8th grade only, the OLS1 estimate is significantly negative, while other four possibly biased models all offer significant positive effects, which are also smaller than the effects of other two tracking status. In short, these five possibly biased models give the impression that being tracked to high ability groups in both 8th and 9th grade or in 9th grade only would benefit students in their 9th grade achievement significantly. The effect of being assigned to high ability groups in 8th grade only, however, is not clear. It could be positive yet less beneficial in comparison with other two tracking status. It could even be detrimental to students' achievement in the following year.

Models that include student's prior ability as well as covariates at class and school level in general offer estimated effects of tracking which are about one half of the size of the counterpart models. In the case of OLS models, the difference between OLS2 and OLS3 is that the latter also include not only student's prior ability but also class-level and school-level covariates. Both OLS2 and OLS3 indicate that students would benefit from being assigned to high ability groups in both 8th and 9th grade or just in 9th grade and the positive effects of these two conditions are not that different. OLS3 estimates the effect of being assigned to high ability groups in both 8th and 9th grade is 2.525, which is about 1/5 of the standard deviation of the 9th grade ability score, and the effect of being assigned to high ability groups in 9th grade only is 2.044. The same model offers a negative but not statistically significant estimate of being assigned to high ability classes in 8th grade only. OLS2 offers slightly smaller but very similar estimates of these tracking statuses. In short, the inclusion of class and school level covariates does not change significantly the estimates of tracking status. In comparison with the OLS1 results, the results show that estimation of tracking effects may be more serious biased by omitting student's prior ability than omitting class and school level covariates.

In the case of PSM models, PSM1 and PSM2 use propensity score estimated by psmatch2 (Leuven & Sianesi, 2003), a user written Stata command, to perform PSM analysis. The results of these two models, which differ by the number of matches selected for each treated case (students being tracked), produce fairly similar results, which are also close to results offered by OLS2 and OLS3. PSM1 and PSM2 assume that by conditioning on observable covariates, the treated and matched untreated cases would differ only in their treatment status (Morgan & Winship, 2007). This assumption would be violated if conditioning on observable covariates still could not remove the impact of unobservable on the treatment selection (Rosenbaum, 2002). The estimations offered by PSM-FE1, PSM-FE2, PSM-FE3, and PSM-FE4 as reported in Table 3 are possible solutions to this problem. These four PSM models attempt to deal with the unobservable either by using fixed effect models to control for selection bias at the school level first and then assume that PSM using propensity score of being tracked estimated by these models can successfully handle the issue of unobservable at both student and school levels (the first approach), or by using PSM to find comparable tracked and untracked students with matching variables at both individual and school level first and then using fixed effects regression to estimate the differences in outcomes between these comparable groups (the second approach).

Table 3 shows that the gains of being tracking to high ability groups as estimated by two models using the first approach, PSM-FE1 and PSM-FE2 are similar to the estimation of OLS2, OLS3, PSM1, or PSM2. The only obvious difference is that PSM-FE1 and PSM-FE2 find gains of being assigned to high-ability groups in 8th grade only are significantly negative, while the gains estimated by OLS2, OLS3, PSM1, and PSM2 are not statistically significant. Gains estimated by PSM-FE3 and PSM-FE4, models using the second approach, on the other hand, are similar to FE1, the model of fixed effects regression at the student level. The estimates of FE1, PSM-FE3, and PSM-FE4 all show significant large gains of students being tracked to high-ability groups in both 8th and 9th grade, a slightly less but also considerable gains of being tracked to high ability groups in 9th grade only, and still smaller but significantly positive gains of being tracked to high ability groups in 8th grade only.

In summary, various PSM models offer fairly similar results regarding the positive impact of being assigned to high ability classes in both 8th and 9th grade or in 9th grade only on 9th grade academic achievement. Two PSM models which combined with propensity scores estimated by fixed effects model predicted the impact of being assigned to high ability classes in just 8th grade could be detrimental to students' academic performance in 9th grade, which is different from the non-significant impact predicted by other four PSM models, which either give non-significant effects or positive effects of being in the same tracking status.

5. Conclusion and Discussion

Does tracking have positive influence on junior high students’ academic performance in Taiwan? Hallinan (1994) has suggested, among other things, that grouping students’ strictly on objective academic criteria may counterbalance the negative effect of tracking. Gamoran (2010) also maintains that high-stakes exams are crucial, and tracking is positive to students’ performance in schools or countries with high-stakes exams. Findings of the present research are consistent with these observations. This study finds that junior high students assigned to high-ability classes mainly based on their motivation and achievement and if these students stayed or selected in high-ability classes in 9th grade, the year of taking high-stake entrance examination, tracking would be positive to their academic performance.

However, the study also finds that the effect of tracking could be negative or non-significant, if students only selected into high ability classes in 8th grade and exited the type of classes in 9th grade. However, the effect of this tracking status is not conclusive since the results of fixed effects regression also show considerable positive gains.

What are possible mechanisms that make tracking to high ability classes be positive to students' achievement in Taiwan? This is an important question that definitely needs further investigation. Using 2003 TEPS data of 9th graders, I explored some possibilities. The preliminary findings show that students claimed to be assigned to high ability groups in 9th grade tended to perceive more teachers in their schools who would praise them to be hard learners, listen to what they thought, concern about their learning after school, have good relationship with them, and could deliver lectures clearly. These students also attended more hours of afterschool learning offered by schools. Their homeroom teachers also perceive these students to be better than other students taught by them. The homeroom teachers were also more demanding in students' academic performance. They also tended to finish teaching 9th grade courses before the second semester started. In short, students of high-ability groups tend to perceive their learning environment as encouraging and friendly. Their teachers held high expectation of them and provided them with an accelerated curriculum. These preliminary findings are also consistent with possible mechanisms that make tracking positive to students learning (Mulkey, et al., 2009; Schofield, 2010). Of course, these perceptions of students of high ability classes may also be outcomes of being tracked to high-ability groups (Ansalone, 2009).

There are other questions need to be investigated in the future. First of all, what are effects of being tracked to middle/low ability groups on students' academic performance? Even though TEPS data cannot answer this question with a similar research design developed by the present study, it does not mean that this is not an important question. An earlier exploration of the effect of tracking on math performance with the same TEPS data, which the present research is its extension, indicated that the effect of middle/low tracks could be positive to those students whose prior math ability is at the lowest stratum. However, if students with high prior math ability were wrongly assigned to middle or low ability groups, these students would suffer the most in terms of their 9th grade math ability score (Chen & Kuan, 2008).

Second, the current research uses students' self-reports of assignment to different tracks to evaluate the effects of tracking. Whether more objective way of measuring tracking, such as using course-based indicators (Lucas & Gamoran, 2002) or possible alternatives offered by TEPS data, may produce different results should also be explored in the future.

Third, this study has demonstrated fairly clearly the influence of possible selection bias in estimating the effects of tracking. It is also clear that different estimation methods attempting to control for unobservable at both individual and cluster level may yield quite different estimates. Which model is the least biased and what other possible methods that can deal with the issue of multilevel endogeneity are questions that need to addressed in the future.

Table 1 Summary statistics of variables used in the analysis (N=12,513)

Variable	Mean	SE	Variable	Mean	SE
9th grade Ability score	58.163	.138	Family structure		
Tracking status			Living with both parents	.809	.005
Never	.338	.006	Single father	.049	.003
8th & 9th	.052	.003	Single mother	.073	.003
8th only	.043	.003	Other	.070	.003
9th only	.166	.005	Own educational expectation		
Male	.504	.006	High school	.153	.005
Ethnicity			Junior college	.200	.005
Minnan	.717	.005	4-year college	.219	.005
Hakka	.139	.004	Graduate school	.226	.005
Mainlander	.104	.004	Other	.203	.005
Aborigine	.016	.001	Parental educational		
Other	.024	.002	expectation		
Parental Education			High school	.105	.004
Junior high or below	.281	.006	Junior college	.318	.006
Senior high	.419	.006	4-year college	.227	.005
Junior college	.151	.004	Graduate school	.233	.005
4-year college	.088	.003	Other	.116	.004
Graduate school	.030	.002	Hours of attending cram schools		
Other	.031	.002	per week		
Parental occupation			None	.259	.005
Farmer or no skilled	.225	.005	Less than 4	.249	.005
Semi-skilled or service	.228	.005	4 to less than 8	.263	.005
Clerical	.074	.003	8 to less than 12	.137	.004
Semi-professional	.084	.003	More than 12	.092	.004
Professional	.151	.004	7th grade ability score		
Other	.237	.005	(centered by class means)	.191	.110
Family monthly income			7th grade class mean ability score		
Under NT$20,000	.101	.004	(centered by school means)	-.172	.035
NT$20,000~NT$49,999	.422	.006	7th grade class mean ability score		
NT$49,999~NT$99,999	.357	.006	standard deviation	-.172	.035
NT$100,000~NT$149,999	.078	.003	7th grade school mean ability		
NT$150,000~NT$199,999	.023	.002	score	.880	.002
Over NT$200,000	.014	.001	7th grade school mean ability		
			score standard deviation	.913	.001
			Remote schools	.053	.002

Note: The sample mean of each variable is estimated with the sampling weight provided by TEPS and SE is the linearized standard error.

Table 2 Logistic regression of being tracked to the high-ability class in both 8th & 9th grade, 8th grade only, or 9th grade

	Both 8th & 9th vs. never		8th only vs. never		9th only vs. never	
Variable	Odds ratio	SE	Odds ratio	SE	Odds ratio	SE
Male	.735*	.100	1.124	.142	1.032	.079
Ethnicity (reference: Minnan)						
Hakka	1.023	.311	1.007	.219	1.369*	.177
Mainlander	.892	.200	.622*	.140	.991	.135
Aborigine	.555	.281	.690	.192	.699	.176
Other	.582	.262	.994	.362	.786	.212
Parental education (reference: Senior high)						
Junior high or below	.742	.123	.972	.156	.932	.088
Junior college	.749	.145	1.016	.195	.759*	.084
4-year college	.827	.250	.956	.231	.876	.146
Graduate school	.430*	.170	.865	.354	.449**	.134
Other	1.127	.636	1.706	.598	1.344	.444
Parental occupation (reference: Other)						
Farmer or no skilled	1.419	.341	.892	.170	1.183	.135
Semi-skilled or service	1.203	.258	1.230	.255	.994	.110
Clerical	1.151	.369	1.195	.305	1.163	.163
Semi-professional	.964	.252	1.122	.321	.989	.153
Professional	1.072	.291	1.486	.343	1.155	.170
Family monthly income (reference: NT$20,000~NT$49,999)						
Under NT$20,000	1.673*	.368	1.180	.248	1.003	.168
NT$49,999~NT$99,999	.938	.153	.977	.144	.895	.079
NT$100,000~NT$149,999	.814	.334	1.092	.293	.574***	.093
NT$150,000~NT$199,999	.278*	,168	1.205	.669	.919	.313
Over NT$200,000	1.964	1.009	.626	.287	.266***	.096
Family structure (reference: Living with both parents)						
Single father	.319**	.121	1.080	.270	.933	.180
Single mother	.793	.212	.721	.198	.925	.145
Other	.585	.313	1.336	.300	1.259	.288
Own educational expectation (reference: Other)						
High school	.683	.256	1.342	.258	.976	.155
Junior college	1.550*	.328	1.525*	.310	1.422**	.187
4-year college	1.518*	.295	1.058	.229	1.281*	.144
Graduate school	1.398	.312	1.132	.238	1.389**	.163
Parental educational expectation (reference: Other)						
High school	.478	.192	1.077	.274	.877	.155
Junior college	1.195	.319	.769	.155	.898	.114
4-year college	1.550	.425	.895	.191	1.297	.176
Graduate school	1.498	.399	1.091	.254	1.260	.184
Hours of attending cram schools per week (reference: None)						
Less than 4	1.498	.352	1.046	.181	1.274*	.141
4 to less than 8	1.911**	.410	1.509*	.271	1.420***	.149
8 to less than 12	2.491***	.610	1.785**	.344	1.675***	.205
More than 12	2.161**	.586	1.247	.315	1.580**	.236
(Cont.)						

Table 2 (Cont.)

Variable	Both 8^{th} & 9^{th} vs. never		8^{th} only vs. never		9^{th} only vs. never	
	Odds ratio	SE	Odds ratio	SE	Odds ratio	SE
7^{th} grade ability score (centered by class means)	1.130^{***}	.011	$.970^{***}$	.008	1.062^{***}	.006
7^{th} grade class mean ability score (centered by school means)	1.083^{*}	.044	.966	.028	1.035	.026
7^{th} grade class mean ability score standard deviation	.903	.059	.981	.042	1.022	.038
7^{th} grade school mean ability score	$.916^{***}$	.023	$.919^{***}$	.018	$.420^{***}$	.083
7^{th} grade school mean ability score standard deviation	1.645^{***}	.158	1.056	.084	1.139	.083
Remote schools	2.144	.842	.883	.339	$.375^{***}$	.100
Constant	$.039^{*}$	.056	1.581	1.748	$.031^{***}$	.019
N	100,63		10,103		11,379	
Log psudo likelihood	-1926.962		-2046.357		-5046.736	
Ward X^2 (df)	$419.29 (41)^{***}$		$117.37 (41)^{***}$		$365.42 (41)^{***}$	
Psudo R^2	.223		.044		.092	

Note: Each regression model is estimated with sampling weights provided by TEPS.
* P < .05 ** P < .01 *** P < .001

Table 1 Causal effects of being tracked to the high-ability class at both 8th & 9th grade, 8th grade only, or 9th grade only estimated by OLS regression, various random effects, fixed effects, and PSM models

Model	Tracking Status	ATT	S. E.[1]	t	p-value	N[2]	N of being tracked
OLS1: Student level covariates without 7th grade ability score	8th & 9th	6.096	.483	12.63	< .001	12,513	547
	8th only	-2.615	.729	-3.59	< .001	12,513	587
	9th only	4.042	,376	10.76	< .001	12,513	1,863
OLS2: Student level covariates with 7th grade ability score	8th & 9th	2.258	.286	7.89	< .001	12,513	547
	8th only	-.554	.431	-1.29	.200	12,513	587
	9th only	1.874	.211	8.89	< .001	12,513	1,863
OLS3: Both student and school level covariates	8th & 9th	2.525	.299	8.43	< .001	12,513	547
	8th only	-.490	.436	-1.12	.262	12,513	587
	9th only	2.044	.218	8.43	< .001	12,513	1,863
RE1: Random effects (cluster: students)	8th & 9th	10.543	.320	32.95	< .001	20,126/ 10,063	547
	8th only	3.671	.556	6.60	< .001	20,206/ 10,103	587
	9th only	8.255	.298	27.71	< .001	22,758/ 11,379	1,863
RE2: Random effects (cluster: schools)	8th & 9th	7.140	.265	26.94	< .001	20,126/ 279	547
	8th only	2.864	.393	7.29	< .001	20,206/ 279	587
	9th only	5.976	.219	27.33	< .001	22,758/ 279	1,863
FE1: Fixed-effects (cluster: students)	8th & 9th	9.860	.261	37.80	< .001	20,126/ 10,063	547
	8th only	7.863	.434	18.12	< .001	20,206/ 10,103	587
	9th only	9.671	,222	43.50	< .001	22,758/ 11,379	1,863
FE2: Fixed-effects (cluster: schools)	8th & 9th	7.629	.266	28.67	< .001	20,126/ 279	547
	8th only	3.133	.403	7.78	< .001	20,206/ 279	587
	9th only	6.304	.222	28.44	< .001	22,758/ 279	1,863
PSM1: One to one matching with propensity scores estimated by matching variables[3]	8th & 9th	2.960	.636	4.66	< .001	10,036	520
	8th only	.455	.679	.067	.503	10,074	558
	9th only	2.210	.396	5.58	< .001	11,286	1,770
PSM2 : One to five matching with propensity scores estimated by matching variables[3]	8th & 9th	3.233	.527	6.13	< .001	10,036	520
	8th only	-.503	.554	-.91	364	10,074	558
	9th only	2.252	.457	4.93	< .001	11,286	1,770
PSM -FE1: One to one matching with propensity scores estimated by fixed- effects model with school as clusters [3]	8th & 9th	2.836	.655	4.33	< .001	10,036	520
	8th only	-3.092	.705	-4.39	< .001	10,074	558
	9th only	2.791	.339	8.23	< .001	12,286	1,770
PSM-FE2: One to five matching with propensity scores estimated by fixed- effects model with school as clusters [3]	8th & 9th	2.247	.601	3.74	< .001	10,036	520
	8th only	-2 .618	.628	- 4.17	< .001	10,074	558
	9th only	2.902	.315	9.21	< .001	12,286	1,770
PSM-FE3: One to one matching first to restrict sample to common support and then fixed- effects regression at student level	8th & 9th	10.600	.542	19.57	< .001	6,988/ 3,494	346
	8th only	7.735	.333	23.24	< .001	14,788/ 7,394	558
	9th only	9.812	.220	44.67	< .001	18,450/ 9,225	1,433
PSM-FE4: One to five matching first to restrict sample to common support then fixed- effects regression at student level	8th & 9th	9.989	.392	25.50	< .001	7,336/ 3,668	520
	8th only	7.735	.333	23.24	< .001	14,788/ 7,394	558
	9th only	9.656	.192	50.27	< .001	19,124/ 9,562	1,770

1. S. E. indicates the standard error. Standard errors of the OLS, fixed effects, and random effects model were robust standard errors. Standard errors of PSM models were estimated with bootstrapping methods.
2. For OLS models, N is the sample size. For fixed effects and random effects models, the figure before "/" is the number of observations, which were observed repeatedly in two waves and nested within the number of groups

reported after “/”. For PSM models, N is the sample matched on common support. Other than using 1 to 1 or 1 to 5 matching, to improve the precision of the estimation, all PSM models were estimated within a common support region as well as a trimming level of 5% and a caliper of .25 standard deviation of the estimated propensity scores.

References

Allison, P. D. (2009). *Fixed Effects Regression Models*. Thousand Oaks, CA: Sage Publications.

Ansalone, G. (2009). *Exploring Unequal Achievement in the Schools: The Social Construction of Failure*. Lanham, MD: Lexington Books.

Ansalone, G. (2010). Tracking: educational differentiation or defective strategy. Educational Research Quarterly, 34 (2), 3-17.

Ansalone, G. & Biafora, F. (2004). Elementary school teachers' perceptions to the educational structure of tracking. *Education*, 125 (2), 249-259.

Arpino, B. & Mealli, F. (2010). The specification of the propensity score in multilevel observational studies *Computational Statistics and Data Analysis*, 55, 1770-1780.

Biafora, F. & Ansalone, G. (2008). Perceptions and attitudes of school principals towards school tracking: Structural considerations of personal beliefs. *Education*, 128 (4), 588-602.

Broaded, C. M. (1997). The Limits and Possibilities of Tracking: Some Evidence from Taiwan. *Sociology of Education*, 70 (1), 36-53.

Buscha, F., Maurel, A., Page, L., & Speckesser, S. (2012). The Effect of Employment while in High School on Educational Attainment: A Conditional Difference-in-Difference Approach. *Oxford Bulletin of Economics and Statistics*, 74 (3): 380-396.

Chang, L.-Y. (2009). *Taiwan Education Panel Survey: Base Year (2001) Student Data and Parent data [on-site release computer file]*. Center for Survey Research, Academia Sinica. License number: TEPS2A002097.

Chen, W.-L. & Kuan, P.-Y. (2010). The effects of tracking on math performance: Evidence from the junior high students in Taiwan. A paper presented at the Spring Meeting on Social Consequences of Economic Uncertainty: Local and Global Perspectives, the Research Committee on Social Stratification and Mobility (RC28), International Sociological Association, May 9-11, University of Haifa, Israel.

Ebbes, P., Böckenholt, U., & Wedel, M. (2004). Regressor and random-effects dependencies in multilevel models. *Statistica Neerlandica*, 58 (2), 161-178.

Gamoran, A. (1987). The stratification of high school learning opportunities. *Sociology of Education*, 60 (3), 135-155.

Gamoran, A. (1992). The Variable Effects of High School Tracking. *American Sociological Review*, 57(6), 812-828.

Gamoran, A. (2010). Tracking and inequality: New directions for research and practice. In M. Apple, S. J. Ball, & L. A. Gandin (Eds.), *The Routledge International Handbook of the Sociology of Education* (pp. 213-228). London: Routledge.

Gamoran, A., & Berends, M. (1987). The effects of stratification in secondary schools: Synthesis of survey and ethnographic research. *Review of Educational Research* 57 (4): 415–435.

Guo, S. & Fraser, M. W. (2010). *Propensity score analysis: statistical methods and application*. Sage: California.

Hallinan, M. T. (1994). Tracking: From Theory to Practice. *Sociology of Education, 67(2)*, 79-84.

Hanushek, E. A. & Woßmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal, 116*, 63–76.

Huang, Min-Hsiung. (2009). Classroom homogeneity and the distribution of student math performance: A country-level fixed-effects analysis. *Social Science Research, 38(4),* 781 -791.

Kulik, J. A. (2004). Grouping, tracking, and de-tracking. In H. J. Walberg, , A. J. Reynolds, & M. C. Wang, (Eds.), *Can unlike students learn together? Grade retention, tracking, and grouping (pp. 157-182),* Greenwich, Conn. : Information Age Pub.

Kulik, J. A.; Kulik, C. C. (1992). Meta-analytic findings on grouping programs. *Gifted Children Quarterly*, 36 (2), 73–77.

Kim, T., Lee, J., & Lee, Y. (2008). Mixing versus sorting in schooling : Evidence from the equalization policy in South Korea. *Economics of Education Review, 27,* 697–711.

Leuven, E. & Sianesi B. (2003). *Psmatch2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing*. Retrieved April 4, 2009, from http://ideas.repec.org/c/boc/bocode/s432001.html#abstract.

Lin, L.-Y. (2010). *The Mixed-Ability Grouping Policy in Taiwan: Influences on Policy and Practice*. (Unpublished doctoral dissertation). The University of Edinburgh, U. K. Retrieved July 25, 2012, from http://www.era.lib.ed.ac.uk/bitstream/1842/5704/2/Lu2010.pdf.

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

Lucas, S. R. & Gamoran, A. (2002). Tracking and the achievement gap. In J. E.

Chubb & T. Loveless (Eds.), *Bridging the achievement gap* (pp. 171-198). Washington, D. C.: The Brookings Institution.

Morgan, S. L. & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York : Cambridge University Press.

Mulkey, L. M., Catsambis, S., Steelman, L. C., & Hanes-Ramos, M. (2009). In L. J. Saha & A. G. Dworkin (Eds.), *International Handbook of Research on Teachers and Teaching* (pp. 1081-1099). New York: Springer Science + Business Media.

Oakes, J. (1985). *Keeping track.* New Haven: Yale University Press.

Rosenbaum, P. R. (2002). *Observational Studies*. New York: Springer.

Schofield, J. W. (2010). International evidence on ability grouping with curriculum differentiation and the achievement gap in secondary schools. *Teachers College Record*, 112 (5), 1492-1528.

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

Van de Werfhorst, H. G. & Mijs, J. J. B. (2010). Achievement inequality and the institutional structure of educational systems: A comparative perspective. *Annual Review of Sociology*, 36, 407-428.