education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 23 Number 62 July 5th, 2015 ISSN 1068-2341

Do Algorithms Homogenize Students' Achievements in Secondary School Better Than Teachers' Tracking Decisions?

Florian Klapproth
University of Luxembourg
Luxembourg

Citation: Klapproth, F. (2015). Do algorithms homogenize students' achievements in secondary school better than teachers' tracking decisions? *Education Policy Analysis Archives, 23*(62). http://dx.doi.org/10.14507/epaa.v23.2007

Abstract: Two objectives guided this research. First, this study examined how well teachers' tracking decisions contribute to the homogenization of their students' achievements. Second, the study explored whether teachers' tracking decisions would be outperformed in homogenizing the students' achievements by statistical models of tracking decisions. These models were akin to teachers' decisions in that they were based on the same information teachers are supposed to use when making tracking decisions. It was found that the assignments of students to the different tracks made either by teachers or by the models allowed for the homogenization of the students' achievements for both test scores and school marks. Moreover, the models' simulations of tracking decisions were more effective in the homogenization of achievement than were the tracking decisions, if the students assigned to the different tracks were at the center of the achievement distribution. For the remaining students, there was no significant difference found between teachers' tracking decisions and the models' simulations thereof. The reason why algorithms produced more homogeneous groups was assumed to be due to the higher consistency of model decisions compared to teacher decisions.
Keywords: ability grouping; tracking decisions; homogenization; secondary school; teachers; statistical models; Luxembourg

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 2/15/2015
Revisions received: 3/11/2015
Accepted: 3/27/2015

¿Pueden lo Algoritmos Homogeneizar Os Logros de los Estudiantes de Escuela Secundaria Mejor que las Decisiones de Seguimiento de los Docentes?
Resumen: Dos objetivos guían esta investigación. En primer lugar, este estudio examinó qué tan bien las decisiones de seguimiento de los docentes contribuyen a la homogeneización de los logros de los estudiantes. En segundo lugar, el estudio analizó si los modelos estadísticos de localización tuvieron mejores resultados que las decisiones de seguimiento de los profesores para la homogeneización de los logros de los estudiantes. Estos modelos tenían información similar a las que usaban los profesores cuando toman decisiones de seguimiento. Se encontró que las asignaciones de los estudiantes a diferentes grupos de nivel hechas ya sea por los profesores o por los modelos homogeneizaron los logros de los estudiantes tanto en exámenes como calificaciones escolares. Por otra parte, modelos que usaban algoritmos para tomar decisiones de localización fueron más eficaces en la homogeneización de logros si los estudiantes asignados a los diferentes grupos de nivel estaban en el centro en la curva de distribución de logros. Para el resto de los estudiantes, no hubo diferencias significativas entre las decisiones docentes y los modelos. La razón por la cual los modelos que usaron algoritmos produjeron grupos más homogéneos fue dada la mayor consistencia en los modelos de decisiones en comparación con las decisiones de los maestros.
Palabras clave: grupos de nivel; seguimiento de decisiones; homogeneización; escuela secundaria; profesores; modelos estadísticos; Luxemburgo

Podem Algoritmos Padronizar os Logros dos Estudantes do Ensino Médio de Acompanhamento Melhor que as Decisões dos Professores?
Resumo: Dois objetivos norteiam esta pesquisa. Em primeiro lugar, este estudo analisou quão bem as decisões dos professores contribuem para a homogeneização desempenho dos alunos. Em segundo lugar, o estudo examinou se modelos estatísticos superaram as decisões dos professores para conseguir a homogeneização do desempenho dos alunos. Estes modelos tinham informações semelhantes aos utilizados pelos professores na tomada de decisões. Verificou-se que a localização dos alunos em diferentes grupos de nível feitas por professores ou por modelos de desempenho homogeneizaram os resultados dos alunos em as provas e as notas. Além disso, os modelos que usaram algoritmos para decisões de localização foram mais eficazes na homogeneização da logros se os alunos estavam no nível médio da curva de distribuição de logros. Para outros alunos, não houve diferenças significativas. A razão pela qual os modelos de algoritmos usados resultaram em grupos mais homogêneos se deve a maior consistência dos modelos de decisão em comparação com as decisões dos professores.
Palavras-chave: grupos de nível; decisões de nível; homogeneização; ensino médio; professores; modelos estatísticos; Luxemburgo

Introduction

Do algorithms assign students to different courses more effectively than teachers, when homogenization of achievement is desired? This article presents results of a study, which strongly affirm this question.

The article is divided into three sections. The first section presents the theoretical background of the study according to different programs for homogenization of achievement in school and with respect to the teacher as a professional who is prone to making inconsistent judgments and decisions. The second section describes the methodology of the study at hand, including the measures of homogeneity that were used for the purpose of the study, and the

algorithms that were derived in order to assign students to different courses. In the third section the results were presented and finally discussed.

I want to make clear preferably at the beginning of this article that this study will not put forward arguments in favor of or against homogenization of students regarding their achievements. It only will be shown that if homogeneity of achievements is a pedagogical goal, one should contemplate the use of (mechanical) algorithms instead of mere human judgment.

Grouping Students Into Different Tracks in Secondary School

In many countries in Europe and beyond, the career of students is mainly determined by the school track they attend in secondary education. For example, students attending and successfully finishing the highest school track will be given the opportunity to go to a college or a university, whereas those who attend one of the lower tracks are usually denied access to higher education. Therefore, one of the most far-reaching decisions affecting students' educational career in these school systems is related to grouping them into different tracks in secondary school.

A major purpose of tracking is to homogenize classroom or track placements in terms of students' personal qualities, performances, or aspirations (Oakes, 1987; Rosenbaum, 1976). With homogenized classes, courses, or tracks, it is commonly assumed to facilitate "didactic fit", i.e., adjustment of learning pace, learning materials, and method of instruction to student ability and concerns (Dar & Resh, 1997).

Tracking, which is the ability-based assignment of students to different secondary-school tracks, is an example of the broader concept of ability grouping in school. For almost a century, ability grouping has been one of the most controversial issues in education. Arguments put forward in favor of ability grouping were in essence that grouping would allow teachers to adapt instruction to the needs of their students, with the possibility to provide high achievers with difficult stuff, and low achievers with rather simple material (cf. Slavin, 1990). In contrast, opponents of ability grouping argue that it is especially bad for the low achievers since they experience a slower pace and a lower quality of instruction (e. g., Gamoran, 1989).

The objective of ability-grouping or tracking has been described as the stimulation of an improvement in regard to school achievement by more individualized and adapted educational methods (Slavin, 1987). Furthermore, educating a class where students have a similar achievement level has been seen as more efficient and less demanding for the teacher than educating a class with students with very heterogeneous achievement levels (Hallinan, 1994).

Research on effects of ability grouping has generated equivocal results, as has been shown in comprehensive reviews from Kulik and Kulik (1987), Slavin (1990), and recently Hattie (2009). Whereas some researchers stress the strength of grouping for high-ability students (e. g., Fuligni, Eccles, & Barber, 1995), others found only small or even negative effects on academic achievement for both high-achievers and low-achievers (Gamoran, 1992; Slavin, 1993).

In the United States or the United Kingdom, tracking is mainly practiced as grouping of students at the class or course level, while students stay in the same school. In school systems with hierarchical tracks, as they are common in some European countries (e.g., Germany, Luxembourg, Switzerland, Austria), but also in Korea, China, Brazil, Russia, and Japan, tracking does take place at the school level. In these school systems, students are allocated by teachers to different schools with different curricula and different final degrees on the basis of their achievements and interests in primary school. Although changing school-tracks in hierarchical systems is possible, it occurs quite rarely (e.g., Baumert, Trautwein & Artelt, 2003; Bellenberg, Hovestadt, & Klemm, 2004; Klapproth, Schaltz, & Glock, 2014).

A recommendation or judgment made by educators, which guides the orientation towards a certain track, predates the actual tracking process, and is, like all human judgments, prone to error.

Attempts to reduce errors in human judgments come – among others – from medical and mental health diagnosis (cf. Grove & Meehl, 1996) where human judgments were replaced by outcomes of statistical models. However, the use of models for judgment or decision-making is quite scarce in the educational practice.

Tracking Decisions as an Example of "Clinical Judgment"

Given that tracking decisions are based on knowledge about students' performance and inferred abilities, homogenizing school achievements through tracking decisions is an example of what has been called "clinical judgment" (Meehl, 1954). Clinical judgments or decisions are rather subjective and based on informal contemplation. In contrast, "mechanical judgment" involves a formal, algorithmic procedure to make a decision (e.g., Grove & Meehl, 1996). Mechanical decisions are often derived from models that mimic human decisions. These models entail some variables and rules about how to combine them. These rules apply "automatically", that is, without intervention of a human decision-maker. In the 1970s, Dawes and colleagues (e.g. Dawes & Corrigan, 1974) showed with various variables that the correlation between the output of a model and a criterion is often higher than the correlation between the decision maker's judgment and the criterion, even though the model is based on the behavior of the decision maker. Up to now, numerous studies have indicated that mechanical decisions outperform clinical decisions in a variety of domains, like medicine (e.g., Clarke, 1985), mental health (e.g., Goldberg, 1969), and education (e.g., Dawes, 1971). Once developed, the application of mechanical decisions requires no expert (e.g., teacher) judgment. Karelaia and Hogarth (2008) reported from a meta-analysis of more than 80 studies published between 1954 and 2007 that the coefficients of correlations between decisions models and external criteria were higher on average by .10 than the correlations between human decisions and the same criteria.

The reason for the superiority of mechanical over clinical decisions was supposed to be predominantly up to unreliability of human decisions (Grove & Meehl, 1996). Even if judges reach decisions by weighting single cues, their weighting is usually inconsistent over time, thus leading to differences in decisions due to variations in weights. Therefore, one might speculate that tracking decisions would have less power for homogenization than an algorithmic combination of students' attributes. This argument implies that teachers make random errors in their decisions. However, it is important to separate these random errors from another form of error, namely bias in teacher decisions, or systematic error. A large body of research indicates that teachers may not only make random errors in their decisions, but they make also systematic errors (e.g., Jussim, 1989; Podell & Soodak, 1993). Mechanical decisions might decrease random errors, but they will still be prone to systematic errors if the variables used for the models introduce a source of biased decisions.

Since virtually all studies concerned with the examination of predictors of tracking decisions have used variants of linear regression analyses, the predominant models of tracking decisions are regression models. In linear regression, the variation of a criterion is explained by the variation of one or more predictors, without necessarily implying that there is a causal relationship between the predictor(s) and the criterion. The amount to which each predictor contributes to the variation of the criterion is expressed by regression weights.

The Present Study

The present study extends a previous study (Kovacs, 2013) using the same sample of 6th-grade students in Luxembourg. In Luxembourg, tracking decisions are made by a council at the end of primary school in 6th grade. This council is composed of primary-school teachers, secondary-school teachers, and school inspectors. Students are oriented to one of two major tracks that constitute the Luxembourgish secondary school (starting at grade 7), which can be described as an

academic track and a vocational track, with each track serving a unique curriculum. The tracks are strictly separated and often located in different schools.

The first aim of this research was to investigate how well teachers' tracking decisions contribute to the homogenization of their students' achievements.

The second aim of the present study was to examine whether teachers' tracking decisions would be outperformed in homogenizing their students' achievements by statistical models of tracking decisions. These models were akin to teachers' decisions in that they were based on the same information teachers are supposed to use when making tracking decisions. The models were varied in regard to the weights the information was given in the models. Whereas one model was an optimal weight regression model (OWRM) where the weighting parameters were estimated by minimizing the prediction error (represented as the sum of squared differences between the observed and the predicted data points), the other model (EWRM) was a simplification of the OWRM, as this model did use equal weights for all predictors involved. With the latter model it was examined whether even in case of an oversimplified weighting of information the model would still assign students to more homogeneous tracks than teachers would do.

When teachers make their tracking decisions, it should be quite easy for them to assign students to the lower track who are at the lower end of the achievement distribution, and to assign students to the higher track who are at the higher end of the distribution. However, students who show achievement scores that are near the decision criterion should require more thorough inspection of their achievements and might also be more likely to be assigned to the "wrong" track.

The following rationale shall illustrate this. Suppose that teachers make decisions about students in a similar way as the models do that are construed to simulate the teachers' decisions. Then, both models and teachers would combine student attributes as a weighted linear function. For example, they might base their decisions on school marks of the main subjects, and might link each school mark with a certain weight. The difference between the models' weighting and the teachers' weighting would be that models keep their weighting constant for all students to be judged, whereas teachers should (presumably unconsciously and on a random basis) vary their weights from student to student (Grove & Meehl, 1996). Due to this variation, the corresponding decision outcomes of the teachers would also vary. If the student to be judged is a low or a high achiever, variations of the weights should alter the numerical outcome of the decision, yet – as long as the outcome is clear beyond the decision criterion – the entire judgment of the teacher would not be altered. More concretely, if a low achiever shows school marks that are far below the class average, variation of the weights would not make a huge difference, so that this student is very likely to be allocated at the lower track. However, if the student shows school marks that are near the decision criterion, variations of the weights would have a much stronger impact since a higher weight might result in a decision for the higher track, and a lower weight for the lower track, independently of the achievement of the student. In contrast, since the models' weights are constant, models of tracking decisions would make the same clear-cut decision for each student independently of his or her placement on the achievement continuum, and would assign all students with equal school marks to the same track.

Since in Luxembourg the numbers of students allocated to either track are roughly the same, the decision criterion for teachers is likely to be located at the center of the achievement distribution. It was therefore examined whether the models' tracking decisions would outperform teachers' tracking decisions for students of two different areas of the achievement distribution, which were the center and the extremes of the distributions.

With Hypothesis 1 it was assumed that the achievement scores of the students would be more homogeneous, that is, more similar to each other, when the students were grouped into different tracks, than when the students were ungrouped. This hypothesis might sound trivial at first

glance, since it appears to be obvious that grouping students according to their achievement would necessarily lead to a decrease of achievement heterogeneity. Yet, suppose that the teachers use much more information for their assignments than mere achievement data, and that these non-achievement data are strongly weighted, than it would be possible that students who perform well could be assigned to the lower track, and students performing much worse could be assigned to the higher track.

Additionally, according to Hypothesis 2, the models' assignments of students to different tracks should be superior in homogenizing the students compared to teachers' assignments, if the achievement of the students was average. However, if the students were low or high achievers, both models and teachers should perform equally well in homogenizing their students' achievements.

Method

The Participants

This research was part of the project "Predictive validity of school placement decisions of primary-school teachers in Luxembourg", funded by grant from the Luxembourgish Fonds National de la Recherche. The data analyzed in this study were provided by the Luxembourgish Ministry of Education (Ministére de l'Education Nationale et de la Formation Professionnelle) and by the Luxembourgish school monitoring. The data set used included data from N = 2,825 students who attended grade 6 in the Luxembourgish school system in school year 2008/2009. These students were a representative sample of an age-cohort of 3,204 students. Correlation analyses revealed that there was only a loose relationship between students who were part of this study and those who were not with respect to the variables used for the models, all rs $\leq$ .06.

51.3 % of the students were girls and 48.7 % boys. Their mean age was 12.52 years (SD = 0.52) at the end of 6th grade in primary school.

Unfortunately, the data did not allow for identifying the different councils making track recommendations, nor the teachers involved. Therefore, we could not account for differences in judgments due to differences between teachers, and we were not able to provide demographic data on the teachers.

Measures of Homogenization of Academic Achievement (Dependent Variables)

The tracking decisions – either made by teachers' or simulated by algorithms derived from regression models – resulted in two groups of students, with each group corresponding to one track. Whether or not the students were more similar to each other in regard to their achievements in the assigned tracks, compared to the entire ungrouped sample, was examined by the variance of achievement data as a measure of homogeneity.

After homogenization, the variance of achievement should be smaller than before homogenization. That is, after the assignment of students to the two tracks, the sum of variances of achievement of the students across the tracks should be smaller than the variance of achievement of all students prior the their assignment to different tracks. Differences in variances can be tested for significance by using the Bartlett test (Bartlett, 1954), which tests the null hypothesis that all k population variances are equal against the alternative that at least two are different. The Bartlett test is robust against different sample sizes, but sensitive in regard to deviations from normality of the distributions.

Besides considering the variances, homogenization of students' achievements was assessed by the degree of overlap that the distributions of both tracks share with each other. In case of perfect homogenization, all high achievers would be assigned to one track, and all low achievers would be assigned to the other track, with no low achievers occurring at the high achievers track,

and vice versa. However, this perfect segregation of students with respect to their achievements is hardly realistic since especially students with average achievements are more or less equally likely to be assigned to either track. Therefore, an overlap of the achievement distributions is likely to occur, and the degree of overlap might serve as an indicator of the success of homogenization. If the achievement distributions of the students of both tracks would overlap only marginally, then the homogenization would be better than if the distributions share a lot of achievement scores.

According to Inman and Bradley (1989), the overlap (*OVL*) of the achievement distributions of both tracks is estimated by

$$OVL = \Phi\left(\frac{x_1 - \mu_1}{\sigma_1}\right) + \Phi\left(\frac{x_2 - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{x_1 - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{x_2 - \mu_1}{\sigma_1}\right) + 1, \qquad (1)$$

$$\text{with } x_{1,2} = -\frac{p}{2} \pm \sqrt{\left(\frac{p}{2}\right)^2 - q}, \qquad (2)$$

$$p = \frac{2\left(\mu_2\sigma_1^2 - \mu_1\sigma_2^2\right)}{\sigma_2^2 - \sigma_1^2}, \qquad (3)$$

$$q = \frac{\mu_1^2\sigma_2^2 - \mu_2^2\sigma_1^2 - 2\ln\left(\sigma_2/\sigma_1\right)\sigma_1^2\sigma_2^2}{\sigma_2^2 - \sigma_1^2}, \qquad (4)$$

and the cumulative standard normal distribution function represented by **Φ**. The *OVL* coefficient indicates the area which one distribution shares with the other distribution. The number of students *n* who are captured by this overlap is given by $n = OVL \times N$.

The means (μ_i) and the variances (σ_i^2) necessary for estimating the overlap were derived from two indicators of the students' academic achievement. Firstly, the school marks of the students obtained in 6th grade in the subjects mathematics, German, and French were used as an indicator of academic achievement. Secondly, test scores were used that were obtained from standardized achievement tests administered in 6th grade, which comprised tasks from the curricular fields mathematics, German, and French. From both test scores and school marks, means and variances were calculated and inserted into the formula for estimating the overlap separately for test scores and school marks.

Predictors (Independent Variables)

Assigning students to different tracks should result in more homogeneous student groups. This assignment was done in reality by Luxembourgish teachers' tracking decisions, or it was simulated by two models that resembled the teachers' decisions. Therefore, the kind of tracking decision (made by teachers or models) served as the independent variable.

Teachers' tracking decisions.

For each student, a tracking decision was recorded that was made by teachers organized within the council. The tracking decisions were coded as 1 (favoring the academic track) or 0 (favoring the vocational track).

Models' simulations of tracking decisions.

Each model produced for each student a simulated tracking decision, based on the variables involved and the regression weights calculated. As with teachers' tracking decisions, model "decisions" were either 1 (favoring the academic track) or 0 (favoring the vocational track).

Statistical Models Mimicking Teachers' Tracking Decisions

Two models of tracking decisions were developed that resembled human-made tracking decisions in regard to the information teachers process in order to make the decision. There were two sources of knowledge that provided hints about the way teacher's tracking decisions are made. The first hint stems from legal authorities, which suggest which information teachers should use when deciding on a recommended track. In Luxembourg, these are the students' school marks obtained in the last year of primary school (in 6th grade), especially school marks in the subjects French, German, and mathematics, and scores of a standardized academic achievement test that is administered in 6th grade, assessing students' competencies in French, German, and mathematics (Reding, 2006). The second hint was provided by scientific literature on predictors of tracking decisions. This literature shows that school marks and test scores are the predominant predictors for tracking decisions (e. g., Arnold, Bos, Richert, & Stubbe, 2007; Bos, Voss, Lankes, Schwippert, Thil, & Valtin, 2004; Klapproth, Glock, Krolak-Schwerdt, Martin & Böhmer, 2013).

Because the tracking decision was a binary variable (vocational track versus academic track), the models estimated tracking decisions by using a form of a generalized linear model, which was logistic regression.

The variables that were used in the models as predictors were the 6th grade school marks of the main subjects (German, French, and mathematics) and the test scores obtained from the domains German, French, and mathematics. All predictor variables were z-standardized due to their varying scales prior to being inserted into the regression equation.

Logistic regression uses a transformed linear combination of predictor variables in order to predict the probability that an individual case will belong to one of the two given categories of the criterion variable:

$$P(Y_i = 1) = \frac{e^{(c+w_1x_{i1}+w_2x_{i2}+\ldots+w_kx_{ik})}}{1+e^{(c+w_1x_{i1}+w_2x_{i2}+\ldots+w_kx_{ik})}}, \qquad (5)$$

where $P(Y_i = 1)$ represents the probability that case i will belong to category 1, assuming that the same set of k cues are considered for each case. Every cue's value for case i is indicated by x_i, while the regression weight for that cue is indicated by w, and c represents some constant.

Optimal regression weights were calculated by minimizing the prediction error, represented as the sum of squared differences between the observed and the predicted data points. The cut-off probability value for classifying cases into predicted groups was .50.

A second model was established which ignored the different contributions of each predictor variable to the prediction of the school track. Instead, this model was as simple as possible, as it used only equal weights (all weights were equal to 1) for each predictor variable. In order to calculate a logistic probability prediction based on an equal weighting of predictor variables, the predictor variables were summed. This summed value was then entered as the only predictor into a logistic regression predicting track recommendations.

The mathematical description of the models was as follows. The dependent variable was the probability of being a member of the academic track ($P(Y_i = 1)$). The logistic regression equation for the optimal weight regression model (OWRM) was:

$$P(Y_i = 1) = \frac{e^{\left(-2.54 + 2.12 \times MarksGerman_i + 2.45 \times MarksFrench_i + 0.85 \times MarksMath_i + 1.12 \times TestGerman_i + 1.03 \times TestFrench_i + 1.40 \times TestMath_i\right)}}{1 + e^{\left(-2.54 + 2.12 \times MarksGerman_i + 2.45 \times MarksFrench_i + 0.85 \times MarksMath_i + 1.12 \times TestGerman_i + 1.03 \times TestFrench_i + 1.40 \times TestMath_i\right)}}$$

and for the equal weight regression model (EWRM):

$$P(Y_i = 1) = \frac{e^{\left(MarksGerman_i + MarksFrench_i + MarksMath_i + TestGerman_i + TestFrench_i + TestMath_i\right)}}{1 + e^{\left(MarksGerman_i + MarksFrench_i + MarksMath_i + TestGerman_i + TestFrench_i + TestMath_i\right)}}$$

Note that both school marks and test scores were z-transformed before being inserted into the models.

Variables Used in the Models (Model Input)

The following variables were used to model human track recommendations.

School marks in 6th grade. School marks for the subjects German, French, and mathematics were given as points, ranging from 0 to 60, with points below 30 representing insufficient achievements.

Results of standardized achievement tests. Test scores were obtained from standardized achievement tests that were administered in 6th grade. These tests comprised tasks from the curricular fields mathematics, German, and French. Test scores were standardized such that the population mean was fixed to 0, and the standard deviation was set to 1.

Results

Table 1 displays the correlation between the tracking decisions made by the teachers and those simulated by the models. As can be seen, the optimal weight regression model (OWRM) represented the teachers' tracking decisions more precisely than the equal weight regression model (EWRM). This result indicates that optimal weighting yielded a better fit between the model and the tracking decisions than did (arbitrary) equal weighting.

Table 1
Correlation Between the Teachers' Tracking Decisions and Those Made by the Models

	Teachers' Tracking Decision (1)	Tracking Decision Simulated by Optimal Weight Regression Model (2)	Tracking Decision Simulated by Equal Weight Regression Model (3)
(1)	1	.867***	.777***
(2)		1	.795***

The differences between the models and the tracking decisions were also displayed by the distributions of students on the different tracks. As Table 2 shows, both the teachers as well as the OWRM assigned more students to the vocational track than to the academic track, whereas the EWRM did the reverse.

Table 2
Distribution of Track Recommendations Made by Teachers and the Two Models

Recommendation for	Teachers	OLRM	EWRM
Academic Track	43.9 % (n = 1240)	43.4 % (n = 1226)	54.8 % (n = 1548)
Vocational Track	56.1 % (n = 1585)	56.6 % (n = 1599)	45.2 % (n = 1277)

The achievement measures of all students who were assigned to the vocational track, and of all students who were assigned to the academic track were then used to calculate the variance of the scores as an indicator of homogeneity. Table 3 depicts the results obtained from each model and from the tracking decisions of the teachers.

Table 3
Measures of Homogeneity for the Entire Sample
(A) Test scores

	Mean		Variance		Sum of Variances
	Track V	Track A	Track V	Track A	
Tracking decision	-0.336	0.736	0.236	0.194	0.430
OWRM	-0.343	0.747	0.213	0.175	0.388
EWRM	-0.468	0.623	0.177	0.209	0.386

(B) School marks

	Mean		Variance		Sum of Variances
	Track V	Track A	Track V	Track A	
Tracking decision	41.429	52.207	26.664	9.531	36.195
OWRM	41.271	52.361	24.744	7.789	32.533
EWRM	39.911	51.171	20.951	12.271	33.222

Note. Upper table: means and variances obtained from test scores; lower table: means and variances obtained from school marks. Track V means vocational track, Track A means academic track. OWRM stands for the optimal weight regression model, EWRM stands for the equal weight regression model.

The table shows that the tracking decisions and the OWRM produced very similar means and variances, whereas the variances produced by the EWRM were smaller for the vocational track and larger for the academic track with both test scores and school marks. However, when the variances were summed up across the tracks, both models resulted in more homogeneous achievements compared to the teachers' tracking decisions.

In order to test Hypothesis 1, the variance of the test scores and the variance of the school marks for all students before the grouping was conducted were estimated. For the entire sample (N = 2,825), the mean and the variance of the test scores were $M_{\text{Test}} = 0.136$ and $s^2_{\text{Test}} = 0.501$, and for the school marks $M_{\text{Marks}} = 46.052$ and $s^2_{\text{Marks}} = 47.774$, respectively. Compared to the variances before the grouping (see Table 3), the grouping of the students actually led to a decrease of the variances, independently of whether the grouping was done by the teachers or by statistical models.

To test Hypothesis 1, the Bartlett test was used. With the Bartlett test it was examined whether there was a significant difference between the sum of the variances across the tracks and the variance of the entire sample, separately for each achievement measure. The corresponding null hypothesis stated that all variances were of the same amount. This means that if one of the four

variances was significantly different from any other variance, the Bartlett test would produce a significant value.

The Bartlett test is a Chi-square statistic, which is defined as follows:

$$\chi^2 = \frac{2.303}{C}\left(\sum_i n_i - p\right)\lg\left(s^2\right) - \sum_i \left(n_i - 1\right)\lg\left(s_i^2\right), \qquad (6)$$

$$\text{with } C = 1 + \frac{1}{3\left(p-1\right)} \times \left[\sum_i \frac{1}{n_i - 1} - \frac{1}{\sum_i n_1 - p}\right], \quad (7)$$

with s^2 being the pooled variance of the samples, s_i^2 being the variance within each sample, p being the number of samples compared, and n_i being the size of each sample.

For the test scores as an indicator of homogeneity, Chi-square resulted in χ^2 (df = 3) = 56.182, $p < .001$, indicating that the variance of the entire sample was significantly larger than any other variance. Significant differences between the variances were also obtained for the school marks, χ^2 (df = 3) = 138.761, $p < .001$. Thus, Hypothesis 1 was confirmed since the homogeneity of achievement was substantially increased after the tracking compared to prior to the tracking.

The next step was the assessment of homogenization in different areas of the achievement scores distributions. With both school marks and test scores, the score distributions were divided into four equal-sized parts. After that, the students of the outer parts of the distributions (i.e., the low and the high achievers) were put together to one group, and the remaining students (i.e., the average achievers) formed the second group.

According to Hypothesis 2, the degree of homogenization with students showing average achievements should be stronger when the assignment of the students to the different tracks was made by the models instead of by the tracking decisions. However, if the students were low or high achievers, both models and teachers should perform equally well in homogenizing the students' achievements. To test this hypothesis, the overlap of the distributions as an indicator of homogeneity was assessed by the formula proposed by Inman and Bradley (1989). Table 4 shows the results.

Table 4
Degree of Overlap (OVL) of the Distributions of Achievement Scores

	Low or High Achievers		Average Achievers	
	Test Scores	School Marks	Test Scores	School Marks
Tracking Decision	0.009	0.002	0.462	0.325
OWRM	0.004	0.001	0.373	0.196
EWRM	0.005	0.001	0.332	0.266

Note. OWRM stands for the optimal weight regression model, EWRM stands for the equal weight regression model.

As expected, low and high achievers were placed into the tracks with only a marginal overlap between the achievement distributions, which shows that both the teachers and the models could

easily assign each student to a track that fits her or his academic capabilities. In stark contrast, students showing rather average achievements were classified with a much stronger degree of overlap, which points to the fact that the tracks contained students showing quite diverse achievements, and that the achievements of the students were similar between the tracks.

Differences between the various degrees of overlap were tested for significance by transforming the areas of overlapping distributions into the number of students who were captured by the overlap according to $n = OVL \times N$. The overlap produced by the tracking decisions were compared with the overlap produced by each model, and the models were as well compared with each other, separately for low or high achievers and average achievers. Thus, 12 comparisons resulted in total. The two-proportion z-test was used, which tests against the null hypothesis that the proportions of students covered by the overlap were the same between either the tracking decision and a model's assignment, or between both models' assignments. In order to adjust for alpha cumulation, the significance level was lowered after Bonferroni by factor 3 (resulting in α significance level of $\alpha_{adjusted} = .017$), since three comparison were made per area of achievement (low or high achievers versus average achievers) and per achievement indicator (test scores versus school marks).

There were no significant differences of the overlaps for test scores and school marks between either decision, when the decisions were made for low and high achievers, all ps > .054. However, in case of average achievers, all comparisons produced significant differences. That is, not only were the overlaps significantly smaller when the decisions were made by the models instead of by the teachers (teachers' decisions versus OWRM: $z_{Test} = 4.81$, $p < .001$; $z_{Marks} = 7.80$, $p < .001$; teachers' decisions versus EWRM: $z_{Test} = 7.07$, $p < .001$; $z_{Marks} = 3.42$, $p < .001$), but the models did also differ among each other, with the OWRM being superior for test scores ($z = 2.28$, $p = .011$), and the EWRM being superior for the school marks ($z = -4.42$, $p < .001$).

Discussion

The objective of the present study was twofold. On the one hand, it was examined whether the achievement of students at the end of primary school would be more homogeneous, that is, more similar to each other, when the students were grouped into different tracks, than when the students were ungrouped. This tracking was done both by teachers as well as by statistical models that resembled the teachers' tracking decisions in that they utilized similar information in order to assign the students to different tracks. On the other hand, it was hypothesized that the statistical models would be superior to the teachers in homogenizing the students' achievements after they were assigned to the different tracks, if the students showed rather average achievement. However, both teachers and statistical models should be equally effective in homogenizing the achievement of students who were either on the lower or on the higher end of the achievement continuum.

With respect to the first hypothesis, the assignments of students to the different tracks made either by teachers or by the models allowed for the homogenization of the students' achievements for both test scores and school marks. Compared to the entire sample, the sum of variances of achievement for both tracks were much smaller for both test scores and school marks. Thus, Hypothesis 1 could be confirmed.

Regarding the second hypothesis, it was found that the models' simulations of tracking decisions were more effective in the homogenization of achievement than were the tracking decisions themselves. This, however, was only true if those students were assigned to the different tracks who were at the center of the achievement distribution and therefore supposedly near the decision criterion. For the remaining students, there was no significant difference found between

teachers' tracking decisions and the models' simulations thereof. Hence, Hypothesis 2 was also confirmed.

Since the models differed in the way the achievement information was weighted for the assignment of a student to either the vocational or the academic track, it was no surprise that they differed also in the degree of homogenization. It was found that the equal-weight regression model (EWRM) was superior to the optimal-weight regression model (OWRM) when test scores served as indicators of achievement. However, when homogeneity was measured on the basis of school marks, the OWRM outperformed the EWRM. This difference was presumably due to the fact that in the OWRM school marks had on average larger weights than test scores, whereas in the EWRM all weights were equal, such that the school marks were comparatively more heavily weighted than in the OWRM. Hence, it appears that homogenization of a certain achievement indicator is more effective if this indicator is given more weight in a model than any other indicator.

The results of this study confirm a large body of research which indicated that so-called mechanical judgments usually outperform "clinical" judgments in a broad variety of domains (cf. Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meehl, 1954). Grove and colleagues (Grove et al., 2000) included nine studies in their meta-analyses which were concerned with comparisons of clinical and mechanical predictions in educational contexts, and all of these studies reported an advantage in favor of mechanical judgments. For the present study, it was therefore expected that teachers' inconsistency that might be inherent in making tracking decisions would make models of tracking decisions more accurate than the tracking decisions themselves.

What does the overlap of the achievement-scores distributions obtained from both tracks mean in terms of the students who were captured by the overlap? Students of one track whose achievements fall beyond the intersection of the distributions would show achievement scores that are more similar to the average score of the opposite track than to the average score of their own track. Hence, these students might be termed "misclassified" (cf. Klapproth, Krolak-Schwerdt, Hörstermann, & Martin, 2013) as they contribute more to the heterogenization than to the homogenization of achievements within their track.

Should students be taught in homogenized courses or tracks? Although most experts agree that high-ability students tracked into homogeneous high-ability groups benefit from the tracking, evidence from highly controlled studies has been brought that low-ability students tracked into low-ability groups do not (Argys, Rees, & Brewer, 1996; Duru-Bellat & Mingat, 1998; Hoffer, 1992; Kerckhoff, 1986). Becker and colleagues (Becker, Lüdke, Trautwein, Köller, & Baumert, 2012) investigated the effect of tracking in the German secondary school system and showed that students who attended an academic track achieved higher scores in an intelligence test than students who attended a vocational track, even though prior achievement and intelligence level were controlled. Becker and colleagues (Becker et al., 2012) attributed these differences to the higher educational quality of academic tracks, compared to vocational tracks. Similar results were found by Schaltz and Klapproth (2014) for Luxembourgish secondary schools. However, if these lower tracks were more stimulating, challenging and taught by well-trained teachers, there might me more gains from tracking for these students (Hattie, 2009). Ability-grouping is, however, not restricted to allocate students to different tracks. Another form of creating homogeneous learning groups is within-class grouping, which can be defined as the teacher's practice of forming groups of students of similar ability within an individual class (Hollifield, 1987). In contrast to between-school tracking, within-class grouping has been shown to be much more effective in regard to students' achievements, even for the low-achievers (Kulik & Kulik, 1992). Thus, it seems that homogenization of students' achievements might be beneficial in some instances, provided that learning materials and teaching are appropriately varied according to the ability levels of the students (Hattie, 2009).

Limitations of the Study

Two limitations pertinent to this study can be assumed. The first one is related to the number of regression models that were used to simulate teachers' tracking decisions. Since only two models were applied, it could be argued that these models are only special cases of the whole family of regression models, and it might be the case that different models would produce assignments of students that are inferior to the assignments made by teachers. Certainly, this argument is valid on a general level. However, in this study it was shown that even when a regression model was used that ignored the different weightings of student characteristics which were used to come to a decision about the track a student should be placed in, this model was more effective in homogenizing students' achievements than the teachers were. Hence, it was demonstrated that regression models' "decisions" outperform human-made decisions regardless the weights that were ascribed to a distinct piece of information, and I therefore presume that with this study not only special examples, but a class of regression models was examined with respect to their ability of classifying students.

The second limitation refers to the question of whether or not regression models are valid models of human (teacher) judgment. Using linear equations to model decisions has major theoretical implications. First, the relationship between the predictors and the criterion is assumed to be linear (or log-linear if the criterion is a binary variable); second, a low weight of one predictor can be compensated by a high weight of another predictor, without changing the value of the criterion; third, the criterion is always based on all predictors inserted into the regression model. None of these assumptions is necessarily true, and especially the latter two assumptions have been called into question by research dealing with judgment heuristics. Kahneman and Tversky, for instance, have argued that people often base their decisions on simplified strategies instead of full, systematic analyses of the available data (Kahneman & Tversky, 1973; Tversky & Kahneman, 1974). One hypothesis about how people make decisions beyond taking all available information into account is the take-the-best heuristic, suggested by Gigerenzer and Goldstein (1996). This heuristic is an instance of so-called fast-and-frugal heuristics, which are fast in execution and frugal in the information used (Gigerenzer, 2008).

The take-the-best heuristic has been applied in several studies comparing the effectiveness of simple linear models to that of heuristic models (Dhami & Ayton, 2001; Dhami & Harries, 2001; Hogarth & Karelaia, 2006, 2007; Gigerenzer, 2008; Katsikopoulos, Pachur, Machery, & Wallin, 2008) and has also been applied to predictions of high school dropout rates (Gigerenzer, Todd, & the ABC Research Group, 1999). Consistently, heuristic models outperformed regression models when the sample sizes were rather small and the regression models rather complex.
Taken these arguments and findings into consideration, one might wonder whether the application of a fast and frugal algorithm might even outperform variants of linear regression models in homogenizing students' achievements. Future work may continue here.

Conclusion

This study brought evidence that the ability grouping of students – exemplified as the placement of students to different tracks in secondary school – leads to the homogenization of their achievements. Moreover and more importantly, it was shown that homogenization of students' achievements was more effective if the ability grouping was done by the aid of algorithms instead of by teachers. The algorithms that were used in this study were based on regression analysis and – concerning the information that was used in the algorithms – similar to real-live tracking decisions made by teachers. The reason why algorithms produced more homogeneous groups was simply that

they were more consistent than teachers, when students had to be grouped who were average achievers. Especially for those students, the use of algorithms is recommended.

References

Argys, L., Rees, D., & Brewer, D. (1996). Detracking America's schools: Equity at zero cost? *Journal of Policy Analysis and Management, 15*, 623–645. http://dx.doi.org/10.1002/(SICI)1520-6688(199623)15:4<623::AID-PAM7>3.0.CO;2-J

Arnold, K.-H., Bos, W., Richert, P., & Stubbe, T. C. (2007). Schullaufbahnpräferenzen am Ende der vierten Klassenstufe. [Preferences for student careers at the end of 4th grade.] In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, K. Schwippert & R. Valtin (Hrsg.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 271–297). Münster, Germany: Waxmann.

Bartlett, M. S. (1954). A note on the multiplying factors for various chi-squared approximations. *Journal of the Royal Statistical Society, Series B, 16*, 296-298.

Baumert, J., Trautwein, U., & Artelt, C. (2003). Schulumwelten – institutionelle Bedingungen des Lehrens und Lernens. [Institutional conditions of teaching and learning.] In Deutsches PISA-Konsortium (Ed.), *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 261-331). Opladen, Germany: Leske + Budrich.

Becker, M., Lüdke, O., Trautwein, U., Köller, O., & Baumert, J. (2012). The differential effects of school tracking on psychometric intelligence: do academic-track schools make students smarter? *Journal of Educational Psychology, 104*, 682-699. http://dx.doi.org/10.1037/a0027608

Bellenberg, G., Hovestadt, G., & Klemm, K. (2004). *Selektivität und Durchlässigkeit im allgemein bildenden Schulsystem. Rechtliche Regelungen und Daten unter besonderer Berücksichtigung der Gleichwertigkeit von Abschlüssen.* [Selectivity and permeability in school systems. Legal regulations and data in consideration of equality of certifications.] Duisburg, Germany: Arbeitsgruppe Bildungsforschung/Bildungsplanung Universität Duisburg-Essen.

Bos, W., Voss, A., Lankes, E.-M., Schwippert, K., Thiel, O., & Valtin, R. (2004). Schullaufbahnempfehlungen von Lehrkräften für Kinder am Ende der vierten Jahrgangsstufe. [Tracking decisions from teachers at the end of 4th grade.] In J. Doll & M. Prenzel (Hrsg.), Bildungsqualität von Schule: Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsverbesserung (S.191–228). Münster, Germany: Waxmann.

Clarke, J. R. (1985). A comparison of decision analysis and second opinions for surgical decisions. *Archives of Surgery, 120*, 844-847. http://dx.doi.org/10.1001/archsurg.1985.01390310080018

Dhami, M. K., & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision Making, 14*, 141-168. http://dx.doi.org/10.1002/bdm.371

Dhami, M. K., & Harries, C. (2001). Fast and frugal versus regression models of human judgment. *Thinking & Reasoning, 7*, 5-27. http://dx.doi.org/10.1080/13546780042000019

Dar, Y., & Resh, N. (1997). *Enhancing education in heterogeneous schools: Theory and application*. Ramat-Gan, Israel: Bar-Ilan University Press.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26*, 180-188. http://dx.doi.org/10.1037/h0030868

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95-106. http://dx.doi.org/10.1037/h0037613

Duru-Bellat, M., & Mingat, A. (1998). Importance of ability grouping in french collèges and its impact upon pupils' academic achievement. *Educational Research and Evaluation, 4*, 348–368.

Fuligni, A. J., Eccles, J. S., & Barber, B. L. (1995). The long-term effects of seventh-grade ability

grouping in mathematics. *The Journal of Early Adolescence, 15*, 58-89. http://dx.doi.org/10.1076/edre.4.4.348.6951

Gamoran, A. (1989). Measuring curriculum differentiation. *American Journal of Education, 97*, 129-143. http://dx.doi.org/10.1086/443918

Gamoran, A. (1992). The variable effects of high school tracking. *American Sociological Review, 57*, 812-828. http://dx.doi.org/10.2307/2096125

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science, 3*, 20-29. http://dx.doi.org/10.1111/j.1745-6916.2008.00058.x

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650-669. http://dx.doi.org/10.1037/0033-295X.103.4.650

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Goldberg, L. R. (1969). The search for configural relationships in personality assessment: The diagnosis of psychosis vs. neurosis from the MMPI. *Multivariate Behavioral Research, 4*, 523-536. http://dx.doi.org/10.1207/s15327906mbr0404_7

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*, 293-323. http://dx.doi.org/10.1037/1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19-30. http://dx.doi.org/10.1037/1040-3590.12.1.19

Hallinan, M. (1994). Tracking: From theory to practice. *Sociology of Education, 67*, 79-84. http://dx.doi.org/10.2307/2112697

Hattie, J. (2009), *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Hoffer, T. (1992). Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation and Policy Analysis, 14*, 205-227. http://dx.doi.org/10.3102/01623737014003205

Hogarth, R. M., & Karelaia, N. (2006). "Take-the-best" and other simple strategies: Why and when they work well with binary cues. *Theory and Decision, 61*, 205-249. http://dx.doi.org/10.1007/s11238-006-9000-8

Hollifield, J. (1987). *Ability grouping in elementary schools*. Urbana, IL: ERIC Clearinghouse on Elementary and Early Childhood Education.

Inman, H. F., & Bradley, E. L. (1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics–Theory and Methods, 18*, 3851-3874. http://dx.doi.org/10.1080/03610928908830127

Jussim, L. (1989). Teacher expectations: Self-fulfilling prophecies, perceptual biases, and accuracy. *Journal of Personality and Social Psychology, 57*, 469-480. http://dx.doi.org/10.1037/0022-3514.57.3.469

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237-251. http://dx.doi.org/10.1037/0022-3514.57.3.469

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*, 404-426. http://dx.doi.org/10.1037/0033-2909.134.3.404

Katsikopoulos, K. V., Pachur, T., Machery, E., & Wallin, A. (2008). From Meehl to fast and frugal

heuristics (and back): New insights into how to bridge the clinical-actuarial divide. *Theory & Psychology, 18*, 443-464. http://dx.doi.org/10.1177/0959354308091824

Kerckhoff, A. (1986). Effects of ability grouping in British secondary schools. *American Sociological Review, 51*, 842-858. http://dx.doi.org/10.2307/2095371

Klapproth, F., Glock, S., Krolak-Schwerdt, S., Martin, R., & Böhmer, M. (2013). Prädiktoren der Sekundarschulempfehlung in Luxemburg. Ergebnisse einer Large-Scale-Untersuchung. [Predictors of track recommendations in Luxembourg: Results of a large-scale study.] *Zeitschrift für Erziehungswissenschaft, 16*, 355-379. http://dx.doi.org/10.1007/s11618-013-0340-1

Klapproth, F., Krolak-Schwerdt, S., Hörstermann, T., & Martin, R. (2013). Predictive validity of tracking decisions: Application of a new validation criterion. In M. Spiliopoulou et al. (Eds.), *Data analysis, machine learning, and knowledge discovery. Studies in classification, data analysis, and knowledge organization* (pp. 61-69). Berlin: Springer.

Klapproth, F., Schaltz, P., & Glock, S. (2014). Elterliche Bildungsaspiration und Migrationshintergrund als Prädiktoren für Schulformwechsel in der Sekundarstufe 1: Ergebnisse einer Längsschnittstudie. [Parental aspirations and migration status as predictors of track changes in compulsory secondary school: Results of a longitudinal study.] *Zeitschrift für Erziehungswissenschaft, 17*, 323-343. http://dx.doi.org/10.1007/s11618-014-0536-z

Kovacs, C. (2013). *(How) can formal modeling improve educational achievement judgments?* Dissertation. Luxembourg: Faculty of Language and Literature, Humanities, Arts and Education.

Kulik, J. A., & Kulik, C. L. C. (1987). Effects of ability grouping on student achievement. *Equity and Excellence, 23*, 22-30. http://dx.doi.org/10.1080/1066568870230105

Kulik, J. A., & Kulik, C. L. C. (1992). Meta-analytic findings on grouping programs. *Gifted Child Quarterly, 36*, 73-77. http://dx.doi.org/10.1177/001698629203600204

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press. http://dx.doi.org/10.1037/11281-000

Oakes, J. (1987). Tracking in secondary schools: A contextual perspective. *Educational Psychologist, 22*, 129-153. http://dx.doi.org/10.1207/s15326985ep2202_3

Podell, D. M., & Soodak, L. C. (1993). Teacher efficacy and bias in special education referrals. *Journal of Educational Research, 86*, 247-253. http://dx.doi.org/10.1080/00220671.1993.9941836

Reding P. (2006). *Le passage primaire post-primaire. Analyse de la procédure d'orientation.* [The transition between primary and secondary school. Analysis of the transition decision.] Luxembourg: Ministère de l'Éducation Nationale et de la Formation professionnelle.

Rosenbaum, J. E. (1976). *Making inequality*: *The hidden curriculum of high school tracking.* New York, NY: Wiley.

Schaltz, P., & Klapproth, F. (2014). The effect of ability-based tracking in secondary school on subsequent school achievement. A longitudinal study. *British Journal of Education, Society, & Behavioural Science, 4,* 440-455.

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research, 57*, 293-336. http://dx.doi.org/10.3102/00346543057003293

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research, 60*, 471-499. http://dx.doi.org/10.3102/00346543060003471

Slavin, R. E. (1993). Ability grouping in the middle grades: Achievement effects and alternatives. *The Elementary School Journal, 5*, 535-552. http://dx.doi.org/10.1086/461739

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131. http://dx.doi.org/10.1126/science.185.4157.1124

About the Author

Florian Klapproth
University of Luxembourg. Now at Marburg University, Germany
florian.klapproth@icloud.com
Florian Klapproth has got his Diploma degree in Psychology in 1999 at Goettingen University, Germany, and has made his PhD at Hildesheim University, Germany, in 2003. From 2003 to 2009, he worked as a Research Assistant at the Institute of Psychology and Working Sciences, Technical University of Berlin. During this time, he accomplished a second book, dealing with waiting and time perception. In April 2010 he obtained his postdoctoral lecture qualification (Habilitation) in Psychology. He is now working at Marburg University as a professor. Florian Klapproth is member of Deutsche Gesellschaft fuer Psychologie (German Society for Psychology), European Association for Research on Learning and Instruction, International Society for the Study of Time, and Society for Judgment and Decision Making.

education policy analysis archives

Volume 23 Number 62 July 5th, 2015 ISSN 1068-2341