A Model-Based Approach to Predicting Graduate-Level Performance Using Indicators of Undergraduate-Level Performance

Judith Zimmermann
Department of Computer Science
ETH Zurich, Switzerland
judith.zimmermann@inf.ethz.ch

Kay H. Brodersen
Department of Computer Science
ETH Zurich, Switzerland
kay.brodersen@inf.ethz.ch

Hans R. Heinimann
Department of Environmental Systems Science, ETH Zurich, Switzerland
hans.heinimann@env.ethz.ch

Joachim M. Buhmann
Department of Computer Science
ETH Zurich, Switzerland
jbuhmann@inf.ethz.ch

The graduate admissions process is crucial for controlling the quality of higher education, yet, rules-of-thumb and domain-specific experiences often dominate evidence-based approaches. The goal of the present study is to dissect the predictive power of undergraduate performance indicators and their aggregates. We analyze 81 variables in 171 student records from a Bachelor's and a Master's program in Computer Science and employ state-of-the-art methods suitable for high-dimensional data-settings. We consider regression models in combination with variable selection and variable aggregation embedded in a double-layered cross-validation loop. Moreover, bootstrapping is employed to identify the importance of explanatory variables. Critically, the data is not confounded by an admission-induced selection bias, which allows us to obtain an unbiased estimate of the predictive value of undergraduate-level indicators for subsequent performance at the graduate level. Our results show that undergraduate-level performance can explain 54% of the variance in graduate-level performance. Significantly, we unexpectedly identified the third-year grade point average as the most significant explanatory variable, whose influence exceeds the one of grades earned in challenging first-year courses. Analyzing the structure of the undergraduate program shows that it primarily assesses a single set of student abilities. Finally, our results provide a methodological basis for deriving principled guidelines for admissions committees.

1. Introduction

1.1 Consequences of the Bologna Process

The Bologna process is a large cross-national effort to redesign the European system of higher education. The main objective of this initiative is to establish more comparable, compatible, and coherent higher education systems across Europe. The process was launched with the Bologna declaration in 1999; 15 years later, 47 European countries have signed the declaration and are committed to redesigning their systems. Meanwhile, many universities in continental Europe have adopted two-cycle study programs – Bachelor's and Master's – that facilitate the mobility of students and staff across Europe. The comparability of such programs has led to rankings and league tables, strong competition among universities in recruiting talented students, and, consequently, increasingly selective admissions (Kehm, 2010). In the foreseeable future, the admission of talented international students will also prove vital for universities in Europe because demographic changes will cause the number of domestic students to drop dramatically (Ritzen, 2010). In the field of Computer Science, where the demand is high for a well-qualified workforce, this decline in student enrolment will be particularly critical.

In the 2012 Bologna Process implementation report (EACEA, 2012), Switzerland was one of the few countries classified as *open*, which means it has high outward degree mobility and even higher incoming degree mobility. Indeed, ETH Zurich attracts many international students especially to the Master's program in Computer Science, which is taught in English. In 2013, approximately 50% of the students graduating from this program held an external Bachelor's degree. These rather recent developments highlight the necessity of rigorously organizing admission policies so that students with a high probability of successfully completing a degree program can be effectively selected. The problem, however, is deciding how to pinpoint the best measure of study success and identify respective admissions instruments that are reliable and valid.

1.2 PREDICTING FUTURE ACADEMIC SUCCESS

One of the most common tasks of Educational Data Mining (EDM) is filtering out information that can be used to model a student's performance when predicting future academic success (Baker and Yacef, 2009; Romero and Ventura, 2010). Those investigations cover entire study programs as well as individual courses and tasks, e.g., in intelligent online-learning systems. Different statistical methods and data mining (DM) techniques are used to address this problem, ranging from descriptive statistics and regression analysis to decision trees, neural networks, and Bayesian networks (Peña-Ayala, 2014; Romero and Ventura, 2010). In this paper, the concept of DM is utilized in a broad sense and includes methodologies to detect patterns in data in general. In the following sections, we first discuss the challenge of defining and measuring study success, the target variable. Afterward, we present potential explanatory variables for predicting that target variable and then describe the DM techniques used for making those predictions.

1.2.1 Target variable

Study success is difficult to quantify; no scientifically or colloquially uniform definition exists (Hartnett and Willingham, 1980; Ramseier, 1977). Nevertheless, that term has repeatedly been framed within the literature (Camara, 2005; Hartnett and Willingham, 1980; Oswald et al., 2001; Rindermann and Oubaid, 1999; Willingham, 1974). For example, Oswald et al. (2001) model study success as a twelve-dimensional construct that is subdivided into three areas: intellectual behavior, interpersonal behavior, and intrapersonal behavior. Although such elaborate constructs most likely represent study success better than cruder ones, they are harder to measure. Rindermann and Oubaid (1999) propose the following six simpler measures: completion of studies, grade point average (GPA), study duration, student satisfaction, professional qualifications, and professional success, whereat most studies rely on GPA as the most appropriate measure (Baron-Boldt et al., 1988; Poropat, 2009; Trapmann et al., 2007). However, any feasible measure can only be a proxy for *true* study success by a student.

1.2.2 Explanatory variables

Since the early days of educational research, the transition from high school to college has received much attention (Astin, 1993; Atkinson and Geiser, 2009; Conley, 2005; Fetter, 1997; Willingham et al., 1985). Three indicators are of particular importance when predicting future success during that transitional period: GPA obtained at the secondary school, intelligence quotient (IQ), and self-efficacy (Preckel and Frey, 2004). A meta-analysis by Trapmann et al. (2007) measures correlations (mean corrected validities) between high school GPA and undergraduate grades in the range of 0.26 to 0.53. In addition, Atkinson and Geiser (2009) emphasize that high school grades are the best known predictors of student readiness for undergraduate studies, regardless of the quality and type of high school attended, while a standardized admissions test provides useful supplementary information.

The indicative value of several types of explanatory variables has also been assessed during the transition from undergraduate to graduate studies. The predictiveness of indicators of undergraduate achievements for future graduate-level performance has been shown in many studies, revealing explained variances from 4% to 17% (Agbonlaho and Offor, 2008; Downey et al., 2002; Evans and Wen, 2007; Koys, 2010; Kuncel et al., 2001; Lane et al., 2003; Owens, 2007; Timer and Clauson, 2010; Truell et al., 2006). Another type of graduate admissions instrument is the standardized test, such as the Graduate Record Examination (GRE®) General Test, whose validity has been documented in several investigations (Bridgeman et al., 2009; Kuncel et al., 2001). Other researchers measure enabling factors such as language skills and examine the relationship between language proficiency and future study success (Cho and Bridgeman, 2012; Graham, 1987). Current research also analyzes the indicative value of personality traits and the design of respective admissions instruments (e.g., De Feyter et al., 2012; Wikström et al., 2009).

While all of these types of explanatory variables can signal future graduate-study success, it is conceivable that indicators of previous academic achievements are just as useful as they are in the transition between high school and undergraduate studies. When using grades for prediction one must also consider the validity of examinations, grading schemes, and what those grades actually represent (Kane, 2013).

1.2.3 Grades

Applying a factor analysis on school grades, Langfeldt and Fingerhut (1974) find two components that determine achievement: *ability* and *adaptation to the school system*. This is confirmed by research on norm-referenced and criterion-referenced grades (Thorsen, 2014; Thorsen and Cliffordson, 2012). That second dimension is also identified as *student non-cognitive behavior* (Bowers, 2011), *academic ethic* (Rau and Durand, 2000), or *common grade dimension* (Klapp Lekholm and Cliffordson, 2008). It is related to *non-cognitive constructs* such as motivation, effort, self-efficacy, perseverance, and locus of control. Nevertheless, disagreement has also arisen about the quantification of those constructs (Rau, 2001; Schuman, 2001). Here, we refer to that second dimension as *adaptation to the academic culture*.

Klapp Lekholm and Cliffordson (2008) highlight the significance of influences that *construct-irrelevant factors* might have on grades (see also Baird, 2011; Sommerla, 1976; Suellwold, 1983; Tent, 1969). Frey and Frey-Eiling (2009) determine the following impact factors: attractive appearance, knowledge about previous grades, capacity of students to express themselves, examiner's feelings about the abilities of a student, gender, precision of handwriting and mistakes in writing, and knowledge about the grades of older siblings. Those authors even recommend that one systematically correct the grades of students who generally show negatively rated characteristics. Birkel (1978) assesses whether it matters if a good examinee follows a bad one, or vice versa, and find that *good after bad* leads to even better grades while *bad after good* leads to even worse ones.

Although *adaptation to the academic culture* may explain why grades are better predictors of success than standardized test scores, the influence of construct-irrelevant factors might seriously harm the validity of grades as an admissions instrument. Some construct-irrelevant factors have less of an effect on the undergraduate level than on primary or secondary education. For undergraduates, less-personal relationships are found between students and examiners, written examinations are often standard, and student numbers rather than names are used for identification. However, other factors, such as the order in which examinations are corrected, remain a problem.

1.2.4 DM techniques

While EDM aims at promoting scientific and mathematical rigor in educational research (Baker and Yacef, 2009), concerns are still raised about the methodologies employed in validity studies, particularly when assessing the relationship between test scores and future success (Atkinson and Geiser, 2009). Theobald and Freeman (2014) also claim the need for more rigor and propose using regression methods in intervention studies.

Valuable reviews of advanced methodologies within EDM have been conducted by Romero and Ventura (2010) and Peña-Ayala (2014). Two basic approaches can be taken to predict student performance modelling in EDM: regression, where a continuous target is predicted, and classification, where a categorical target is predicted. Relevant to the work described here, Baker et al. (2011) detect a student's preparedness for future learning by applying linear regression models in combination with forward variable selection following a cross-validation scheme. The best model outperforms Bayesian Knowledge Tracing. Rafferty et al. (2013) predict individual student performance from paired interaction data by relying on lasso regression. Both papers conclude by emphasizing the importance of predicting student performance to permit early intervention, while we regard the admission selection process as one of the earliest intervention possible. Herzog (2006) also mentions the suitability of linear regression models for analyzing relatively small datasets.

Different methods have been evaluated for their degree of effectiveness when selecting variables (Romero et al., 2014). They include the use of expert knowledge (Baker et al., 2011) as well as relying upon Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), linear mixed models, and group lasso (Ra and Rhee, 2014). One powerful tool for assessing selection stability is bootstrapping (Efron and Tibshirani, 1994). However, it is not often applied for that purpose in EDM.

1.3 FOCUS OF THIS STUDY

1.3.1 Graduate-level performance in Computer Science within the European context

We evaluate indicators of undergraduate achievements that might relate to graduate-level performance in Computer Science within the European context. Previous examinations have been more concerned with outcomes and drop-out rates in introductory undergraduate courses (Bergin and Reilly, 2006; Nugent et al., 2006; Ventura, 2005). We are not aware of any research on the transition from undergraduate to graduate work in engineering and natural sciences, where indicative values might be stronger than in other fields, since the strongest relationship between high school grades and undergraduate achievements has been found for the former ones (Trapmann et al, 2007). Moreover, typical European programs in Computer Science differ in character from those in North America: rather fixed 3 years mono-disciplinary Bachelor's programs in Europe as opposed to more flexible 4 years programs that include courses outside the major field in the USA (Scime, 2008). In fact, several authors emphasize the importance of examining the validity of admissions instruments for a specific use (Cronbach, 1971; Kane, 2013; Messick, 1989; Newton, 2012).

1.3.2 Unbiased dataset

When prior studies entail indicators of undergraduate achievements, the results might be distorted by an admissions-induced selection bias (Dawes, 1975). This effect has been observed in data that have been collected from study programs with selective and compensatory admissions rules (Figure 1a). In contrast, our data are free of such a bias, as students automatically advance to the Master's program (Figure 1b), and they have been collected within only one institution, thereby enabling us to conduct an in-depth investigation of the statistical relationship between undergraduate and graduate studies.

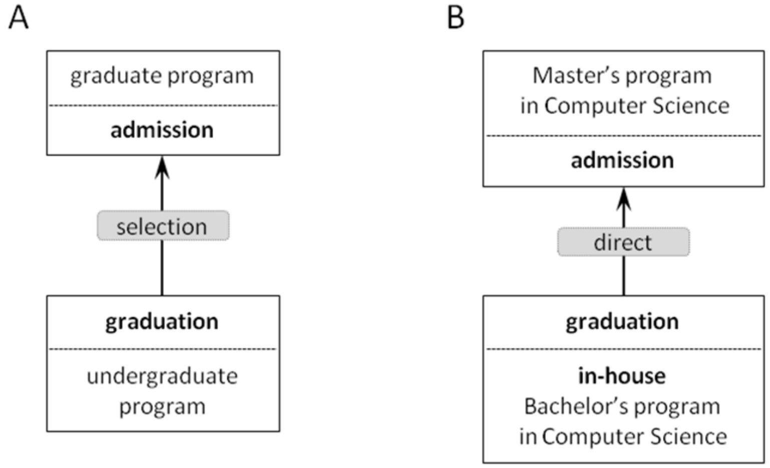

Figure 1: (A) Conventional admissions setting. Graduate programs are typically based on a selective admissions procedure. (B) Admissions setting in this study at ETH. Students who completed the in-house Bachelor's program in Computer Science are directly admitted to the Master's program.

1.3.3 Level of aggregation

When indicators of undergraduate-level course performance are aggregated, the level of aggregation leads to frequent limits on research efforts. In this paper, the spectrum runs from full aggregation (undergraduate grade point average, or UGPA) to none. One can obtain the UGPA by calculating the arithmetic mean across all courses in the undergraduate program, which might average out and hide useful information. The lack of any aggregation strategy provides a set of indicators that might be dominated by construct-irrelevant factors. We hypothesize that an optimal level exists at which undergraduate courses are partially aggregated. For us, this entails three approaches. First, courses should be clustered according to similarity in required abilities and skills. Second, courses should also be clustered in chronological order, because Computer Science programs are typically of a consecutive nature, abilities can develop over time, and students adapt to the academic culture at different paces. Third, in cases where failed examinations are repeated, one might need to cluster the grades achieved in the first attempt and those obtained in the final attempt.

1.3.4 Methodology

We employ linear regression models in combination with different variable-selection techniques. To avoid potentially misleading results caused by methods that are too simplistic (Atkinson and Geiser, 2009), we employ a rigorous DM methodology that includes cross-validation to avoid overfitting, bootstrapping to assess the stability of variable selection, and statistical testing to estimate differences in performance. We compare modern approaches that depend upon adaptive lasso and cross-validated R^2 statistics for performance estimations with those that are more traditional and based on step-wise regression models with AIC and BIC as well as adjusted R^2 statistics. We also adopt an approach to variable selection that relies upon partial correlations, which appears to be particularly appropriate due to anticipated collinearity in the data. With regard to establishing a suitable amount of aggregation, we employ expert knowledge, factor analysis, and the novel minimum transfer-cost principle, which controls the model-order selection process. Our approach elucidates whether results are reliable and sufficiently robust, which is highly valuable in small sample size settings.

1.3.5 Aim

The graduate-admissions process is frequently considered a critically important step in maintaining quality control within higher education. However, rules-of-thumb and domain-specific experiences rather than evidence-based approaches have long dominated university policies (Cuny and Aspray, 2000). Here, we use DM techniques for a thorough investigation of statistical relationships between undergraduate- and graduate-level performances. By combining linear regression models with different methods for variable selection, we aim to i) explore the predictive power of undergraduate indicators and ii) investigate how the meaningful aggregation of grades further improves prediction performance and understanding. We address these questions by analyzing student records from the Bachelor's and Master's programs in Computer Science at ETH Zurich, Switzerland. Our results are also used to derive optimized admissions rules, therewith helping us provide guidance when choosing selection instruments for graduate admission into Computer Science programs and related fields in Europe.

2. DATASET

We analyzed data consisting of 171 student records collected over eight years (2003-2010) from ETH Zurich, Switzerland. Each record comprised 81 variables from a Bachelor's and a Master's program in Computer Science (Tables 1 and 2). Notably, the most challenging and highly selective courses during the first two study years in the Bachelor's program were compulsory for all students. Moreover, all students who completed the ETH Zurich Bachelor's program automatically advanced to the Master's program. Because no selection was conducted at this transition between programs, the data were not confounded by an admissions-induced bias. Therefore, we were able to acquire a complete dataset for all in-house students who graduated from the Master's program.

These data posed two difficulties for the analysis. First, the number of observations was rather small in relation to the number of explanatory variables, which aggravated the risk of overfitting and subsequent overinterpretation. Second, strong collinearities were expected along with the risk of variance inflation in the resulting models.

Table 1: Overview of the Bachelor of Science program in Computer Science at ETH. Numbers in brackets indicate how many courses students must take in each group to fulfill the degree requirements.

First year	**Second year**	**Third year**
Basic courses (9)	*Compulsory courses (11)*	*Core courses (4)*
Calculus	Advanced Computational Science	Algorithms, Probability and Computing
Data Structures and Algorithms	Computer Architecture	Distributed Systems
Digital Design	Computer Networks	Information Security
Discrete Mathematics	Electrical Engineering	Information Systems
Introduction to Programming	Formal Methods and Functional Programming	Modeling and Simulation Visual Computing
Linear Algebra	Information Theory	Software Engineering
Logic	Introduction to Computational Science	*Elective courses (3-4)*
Probability and Statistics	Introduction to Databases	~ 100 courses in various fields of Computer Science
Physics	Operating Systems	
	Software Architecture	*Compulsory major courses (2)*
	Systems Programming	~ 30 courses and seminars
	Theory of Computing	*Compulsory elective courses (3)*
		~ 100 courses in humanities, social and political sciences

Table 2: Overview of the Master of Science program in Computer Science at ETH. Numbers in brackets indicate how many courses students must take to fulfill the degree requirements.

Curriculum		**Focus areas**
Focus courses (4-6)	*Multidisciplinary courses (2-3)*	Computational Science
~ 10 courses in focus area	> 1000 courses offered by 3 universities	Distributes Systems
Elective courses (4-5)		Information Security
~ 100 courses in various fields of Comp. Sci.	*Compulsory elective courses (1)*	Information Systems
	~ 100 courses in humanities, social and political sciences	Software Engineering
Foundations of Comp. Sci. (4)		Theoretical Computer Science
5 courses in total	*Master's thesis*	Visual Computing

The 81 explanatory variables in Table 3 were used to predict the subsequent graduate GPA (GGPA), which we treated as a proxy for graduate-level performance, defining it as the unweighted arithmetic mean of all grades achieved in Master's level courses related to Computer Science. However, the grade earned for the Master's thesis itself was not included because grading schemes varied widely among academic supervisors.

Table 3: Explanatory variables. VN: variable name; italicized font, scalar values; bold type, vectors (lengths given in brackets).

Variable	**VN**	**Scale**	**Comments**
Sex	s	Nominal	Male or female.
Age at registration	a	Ratio	A student's age at the time of enrolment was preferred over alternative measures (e.g., date of birth or age at the time of data acquisition).
Rate of progress	r	Ratio	This variable encoded the number of credits obtained in the Bachelor's program divided by its duration.
Grade achieved in single course	**g** (56)	Interval	These individual explanatory variables included grades achieved in courses from the first and the second year as well as those earned in the third-year core courses. Because the number of examination repetitions fluctuated across students, separate variables were used to capture the grades obtained in the first and final attempts. Unless an examination was repeated, those two variables had identical values. Note that students could not take an examination more than twice. First and final examination attempts differed by 13% for grades achieved in the first year, by 9% for second-year grades, and by 3% for third-year grades. Grades were given on a 6-point scale that included quarter steps (e.g., 5.25), where '6' represented the highest, '1', the lowest, and '4', the minimum passing grade. Because all courses in the first two years were compulsory, no values were available for that part of the data. During the third year, students had some freedom of choice; thus, between 15% and 65% of the values in core courses were missing. Whenever necessary, a random-forest imputation (Breiman, 2001), which can handle cases where up to 80% of the values were absent, was employed to fill in those missing values.
GPA	**gpa** (18)	Interval	Based on the above single-course achievements, several unweighted GPAs, with a precision of two decimal places, were computed for different subsets of courses. These subsets consisted of courses within the entire Bachelor's program, all courses taken during a particular year, and all courses from a particular group (Table 1). Separate variables were computed using first attempts, final attempts, and all attempts. Note that GPAs were calculated before missing values were imputed.
Duration	**t** (4)	Ratio	Three separate explanatory variables were used to capture the time needed to complete each study year; one was used for capturing the time required to finish the entire Bachelor's program.

3. METHODOLOGY

In this paper, the term *model* indicates the use of specific algorithms, either a combination of a variable-selection algorithm and a linear regression model or an adaptive lasso. *Model instance* denotes a fitted model where specific variables are selected and model parameters have been estimated. As our baseline, we use the *null model*, which is a linear regression model with just one parameter to fit the mean of the target variable. The mean squared error (MSE) achieved by the null model equals the variance of the target variable. Thus, the ratio of the difference between the MSE of the predictions obtained by any model and the ones gained by the null model and the MSE of the null model represents R^2 statistics.

After illustrating our data with descriptive statistics, we answered the first research questions on the predictive value of indicators of achievements that are readily available in undergraduate transcripts. Eight different models were run: four models typically used in more traditional educational research, three that we deemed particularly appropriate for the analysis of our data, and one that is rather modern and powerful. To ovoid over-fitting and, therewith, over-estimating prediction performance, we used cross-validation. For estimating the overall prediction performance, we used two layers of cross-validation, one for selecting the best performing model (inner loop) and the other for estimating prediction performance (outer loop). This construct is called the *model-selection framework*.

These eight models were also trained individually on the entire dataset, keeping the inner loop but removing the outer one. This step provided estimations of the *prediction performance of individual models* and respective R^2 statistics for individual models, and enabled us to determine the *best performing models*. The best performing models were then trained on the entire dataset without cross-validation, which led to one model instance each. These instances were analyzed with respect to the selected variables and their individual contribution to the prediction performance. Thereafter, the models were trained individually on 200 bootstrap samples to assess the stability of variable selection.

To answer the second research question about the meaningful aggregation of undergraduate achievements for improving prediction performance, we pursued the following modeling strategy after the scheme of cross-validation. Briefly, we estimated the prediction performance of linear regression models using 10 different sets of explanatory variables that correspond to various levels of aggregation. We employed expert knowledge and factor analysis (FA) for aggregating the variables partially. Notably, feature construction using FA was performed within the cross-validation loop in order to prevent any information leaking from training to test data. By comparing the estimates, we could determine the best performing set of aggregated explanatory variables and, therewith, the best aggregation strategy. In the next step, we assessed the importance of individual explanatory variables within this set. To understand the results better, we conducted a post-hoc investigation on the latent structure of the Bachelor's program. To do so, we employed a novel technique – the minimum transfer-cost principle – to determine the numbers of generalizable factors in that program.

3.1 DESCRIPTIVE ANALYSIS

We used histograms and scatter plots to illustrate the data. Inter-correlation coefficients were calculated among all explanatory variables for estimating the severity of multicollinearity present in the data.

3.2 Model-based analysis using information available from undergraduate transcripts

3.2.1 Individual models

Competing models. We evaluated the prediction performance of eight competing models (detailed below), which combined variable-selection algorithms with linear regression of the following formula:

$$GGPA_i = \beta_0 + \boldsymbol{\beta_1} \cdot \boldsymbol{g_i} + \boldsymbol{\beta_2} \cdot \boldsymbol{gpa_i} + \boldsymbol{\beta_3} \cdot \boldsymbol{t_i} + \beta_4 \cdot s_i + \beta_5 \cdot a_i + \beta_6 \cdot r_i + \varepsilon_i$$

Here, $\boldsymbol{GGPA}$ was a vector containing the GGPAs of all students, $i = 1, \ldots, n$; the dot product was denoted by '$\cdot$'; $\beta_0, \beta_4, \beta_5$, and β_6 were scalars; $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$, and $\boldsymbol{\beta_3}$ were parameter vectors; ε_i was the error term; and the explanatory variables $(\boldsymbol{g_i}, \boldsymbol{gpa_i}, \boldsymbol{t_i}, s_i, a_i, r_i)$ were those that were readily available from undergraduate transcripts, as described in Table 3.

To decrease the risk of overfitting and reduce the multicollinearity of explanatory variables, we chose models that employed rigorous variable selection. All models, except the adaptive lasso, were trained using a two-step procedure that consisted of variable selection and parameter estimation. Four competing models were obtained by selecting variables using AIC (Akaike, 1974) and BIC (Schwarz, 1978) in forward and backward modes (Guyon and Elisseeff, 2003). This was followed by least-squares fitting of linear regression models. These particular models were chosen because they are typically used in traditional educational research.

Three more models were obtained using partial correlation coefficients in a forward-selection setting. First, we selected the explanatory variable that most closely correlated with the target variable. Then, iteratively, we chose the variable that maximized the partial correlation with the target variable, conditioned on the set of already selected variables. This algorithm was applied three times, setting the number of possible explanatory variables at '1' (model PC1), '2' (PC2), or '3' (PC3). Again, linear regression models were fitted using ordinary least squares. These models were chosen because high multicollinearity was expected in the data and the underlying approach assists the selection of rather uncorrelated explanatory variables, while maximizing information content. We set the numbers of variables to be selected as 1, 2, and 3 because we wanted to obtain rather simple models.

To obtain an additional state-of the art model we used adaptive lasso, which provides simultaneous variable selection and parameter estimation and possesses the so-called oracle property (Zou, 2006). Under certain assumptions, this property indicates that the model's prediction performance will be as accurate as the one of the true underlying model. Moreover, adaptive lasso was shown to perform competitively in high-dimensional data settings, where the number of explanatory variables p exponentially exceeds the number of observations n (Bühlmann, and van de Geer, 2011). Those features render this model an interesting alternative to traditional approaches.

Prediction performance and model selection. To assess the prediction performance of our eight models, we employed a 10-fold cross-validation scheme (Breiman and Spector, 1992). First, the data were shuffled randomly and then split into 10 subsamples. One-tenth of the data was reserved for testing while the models were trained on the remaining data. The models were used to predict the dependent variables of the reserve data. This procedure was repeated 10 times so that we obtained an unbiased prediction $\hat{y}_i$. The squared error $SqE_i = (\hat{y}_i - y_i)^2$ was computed for each model and observation $i = 1 \ldots n$, where y_i denoted the observed value of the target variable and MSE was the arithmetic mean of SqEs. Afterward, we calculated the cross-validated R^2 statistics, $cross\text{-}validated\ R^2 = \frac{\sum_{i=1}^{n} cross\text{-}validated\ SqE_i}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$, where $\bar{y}$ denoted the mean of y_i. For comparisons, we determined the adjusted R^2 statistics that computes the fit of a model over the full dataset penalizing the statistics for the number of explanatory variables included.

To identify the model that performed best, we applied two-sample paired t-tests, Bonferroni-corrected for multiple testing, on the above SqEs. As is customary, we then chose the least-complex model that reproduced sufficiently similar results compared to the model with the best performance in absolute numbers.

Contribution of explanatory variables. To determine which explanatory variables contributed most to the prediction performance of the best models, we trained the latter on the entire set of data. The resultant model instances were then analyzed with regard to which variables were selected as well as according to the importance and significance of those variables in the model instance. This goal was accomplished by calculating standardized β-coefficients and performing an ANOVA.

Finally, we assessed the stability of variable selection and therewith the representativeness of the model instances obtained. This step is important for preventing the interpretation of statistical artefacts. Specifically, the model that performed best as well as those models that presented statistically indistinguishable results were each trained on 200 bootstrap samples. Efron and Tibshirani (1994) propose that this sample number is generally adequate for estimating standard errors in most applications. Each time, a sample having the same size as the original dataset was drawn randomly with replacement from the original dataset. The probability of selection was then calculated for each of our explanatory variables and models. Note that our approach of assessing the stability of variable selection is closely related to the one proposed by Meinshausen and Bühlmann (2010), termed stability selection. Importantly, the approach was shown to be adequate for high-dimensional data. It is based on aggregating the results obtained when variable selection is repeatedly applied on subsamples of the data.

3.2.2 Model-selection framework

To estimate the accuracy with which a prediction was generalized to student data that were not part of the analyzed dataset, we employed the following scheme. An additional 10-fold cross-validation loop was wrapped around the portion pertaining to model selection and performance estimations, as described previously. In each fold of this outer loop, model selection using cross-validation was performed on the training data. Notably, omitting the inner cross-validation loop most likely leads to the selection of an overfitting model, which causes prediction performance to collapse in the outer loop.

The best-performing model was then trained on the entire set of data and the resulting model instance was used to form predictions for the test data of the outer loop. By following this procedure, we could derive a single prediction for each student's graduate-level success and calculate respective SqEs and cross-validated R^2 statistics. Finally, we compared the performance of our model-selection framework with that of the null model and used a two-sample t-test on the SqEs of our model-selection framework and on the SqEs of the null model to assess statistical significance.

3.3 LATENT STRUCTURE OF THE UNDERGRADUATE PROGRAM AND AGGREGATION OF EXPLANATORY VARIABLES

Whereas in the first analysis we exclusively relied on information readily available from undergraduate transcripts, our next investigation pursued the goal of improving prediction performance by linear regression. These linear models were trained and tested on different sets of explanatory variables that were constructed more specifically through variable aggregation. We also concentrated on detecting the piece of the undergraduate program that was most informative with respect to future graduate performance.

3.3.1 Specifically aggregated explanatory variables

To understand the optimal level of course aggregation, we considered three means for averaging individual undergraduate courses: i) no aggregation, i.e., each course provides one explanatory variable; ii) partial aggregation, where grades are averaged across related courses; and iii) full aggregation, i.e., the UGPA. For partial aggregation, we explored two alternative clustering approaches: one based on year-wise clustering (YW) and the other obtained through factor analysis. In this way, we determined five alternative sets of explanatory variables based on single courses (SC), YW, FA clustering, a combination of FA and YW, or the UGPA. To deal with repeated examinations, we applied the different aggregation methods to either the first or the final attempt. Thus, 10 competing sets of explanatory variables were considered (Figure 2).

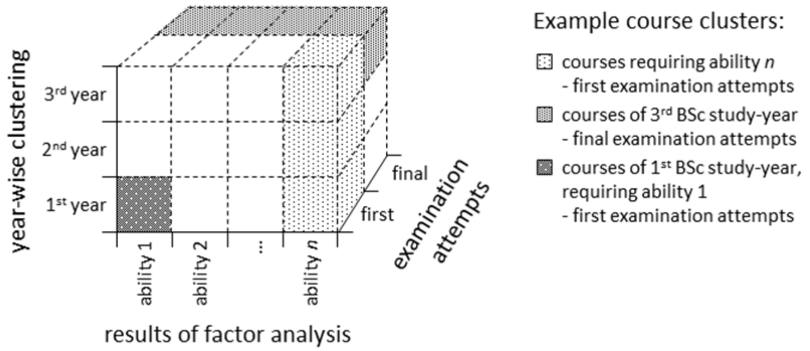
year-wise clustering
3rd year
2nd year
1st year
ability 1
ability 2
...
ability n
results of factor analysis
final
first
examination attempts
Example course clusters:
courses requiring ability n
- first examination attempts
courses of 3rd BSc study-year
- final examination attempts
courses of 1st BSc study-year, requiring ability 1
- first examination attempts

Figure 2: Illustration of three exemplary clustering approaches for partially aggregating the Bachelor's program. Along the x-axis, courses are clustered according to their requirements with respect to student abilities. Along the y-axis, courses are clustered according to the Bachelor's program study year to which they belong. Along the z-axis, courses are split depending upon the examination attempt in which the course grade was achieved.

It is conceivable that the prediction of graduate-level performance might benefit from the use of explanatory variables that represent a student's undergraduate performance in a particular area of scholarship. One way to identify such an area is to form clusters that group together courses with consistent performance levels. To identify this latent structure, we employed an exploratory FA. In particular, we used maximum-likelihood factoring to estimate the parameters within the common-factor model (Costello and Osborne, 2005; Fabrigar et al., 1999; Thurstone, 1947). To determine the number of factors necessary, we performed an χ^2 goodness-of-fit test ($\alpha = 0.01$). Thereafter, we applied varimax rotation to project a so-called simple structure (Fabrigar et al., 1999; Kaiser, 1958). Based on this solution, the factor scores were computed by calculating the average across all courses that had a loading of at least 0.5 on a specific factor (Backhaus et al., 2006; DiStefano, 2009).

The resulting GPAs were used as explanatory variables for predicting the performances of FA and FA plus YW.

To determine which set of explanatory variables generalized best we employed the model-selection procedure presented in Section 3.2.2. On each of the 10 sets of aggregated variables, we trained linear regression models and used them to form predictions according to a 10-fold cross-validation scheme. To avoid over-fitting the data, we performed FA within each cross-validation loop only on the training data. Factor scores were then computed for all students, i.e., for those in the training set and those in the test set. As before, paired t-tests on the SqEs of individual models were used to identify significant differences in prediction performance.

3.3.2 Latent structure of the undergraduate program

In contrast to our expectations, the aggregation of explanatory variables via FA did not significantly improve the accuracy of predictions. This finding suggested that the undergraduate program might not comprise a preeminent factor structure. To examine this possibility, we derived an alternative aggregation of undergraduate courses. The above FA was replaced by the much simpler singular value decomposition (SVD). To test the stability of the SVD for undergraduate grades, we adopted the minimum transfer-cost principle to control the model-order selection process (Frank et al., 2011). First, the data matrix was randomly split in half. The two halves were then aligned to minimize the sum of the entry-wise squared distances. After the SVD was computed on the first half of the matrix, the original matrix was re-constructed. This procedure was repeated for the decompositions of rank 1 through k, where k denoted the number of explanatory variables. The model-order $i = 1 \dots k$ was chosen so that we could minimize the sum of the entry-wise squared distances between the second half of the matrix and the reconstruction of the first half. This allowed us to determine the model-order that generalized best. We note here that the third year was excluded from this analysis because it was incomplete due to missing values; this restriction avoided assumptions implicit in data imputation. Internal consistency of undergraduate grades was assessed using Cronbach's α (Cronbach, 1951).

4. RESULTS AND DISCUSSION

4.1 DESCRIPTIVE ANALYSIS

To estimate the extent of multicollinearity in the data, we computed the inter-correlations between explanatory variables and determined coefficients ranging from $r = 0.04$ to $r = 0.95$. This result demonstrated high multicollinearity in the data and justified our use of the analytical methods outlined previously. Afterward, we assessed the multicollinearity in models more precisely by calculating the variance inflation factor. To illustrate the data, we determined the distributions of the UGPA and GGPA (Figure 3a). For example, at the undergraduate level, students achieved an average GPA of 4.9, with a standard deviation of 0.35. At the graduate level, GPAs were significantly higher, with students earning an average of 5.2, with a comparable standard deviation of 0.36.

Although we see an obvious increase between the grades awarded in the Bachelor's program and those awarded in the Master's program, it is unclear whether one can attribute this to grade inflation or to the tradition of assigning higher grades in graduate courses. The UGPA and GGPA correlate significantly, with $r = 0.65$ (Figure 3b). This statistical dependence justifies why the former is often used to predict the latter. Applying different DM techniques, we next investigate further the importance of the UGPA as well as other indicators of undergraduate performance.

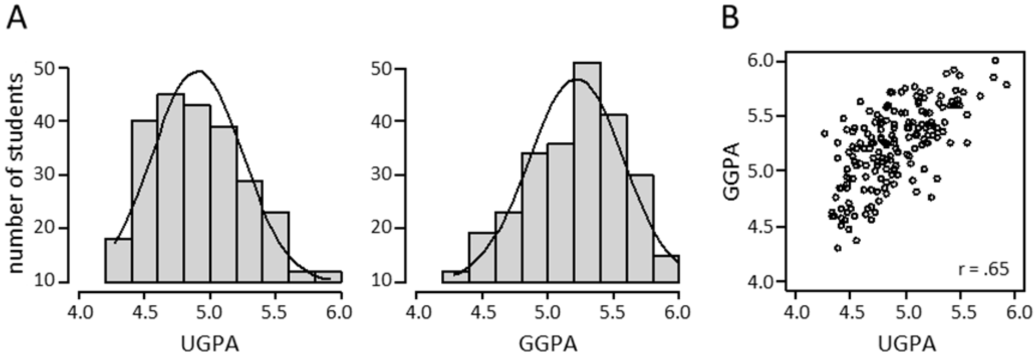

Figure 3: (A) Histograms of UGPA ($\mu = 4.9, \sigma = 0.35$) and GGPA ($\mu = 5.2, \sigma = 0.36$) with normal density function. (B) UGPA versus GGPA. UGPA of individual students is plotted against their GGPA.

4.2 Model-based analysis using information available from undergraduate transcripts

4.2.1 Prediction performance of the model-selection framework

Our first model-based investigation included explanatory variables comprising information that is typically available from undergraduate transcripts. Examples include single grades achieved in an individual course, GPAs of course groups, annual GPAs, or cumulative GPAs. The model-selection framework produced cross-validated R^2 statistics of 0.54, outperforming the null model significantly ($p < 0.001$; two-tailed t-test). Thus, the information available from transcripts explains as much as 54% of the variance in the GGPA. Figure 4a shows the MSEs of the two models (framework and null), where the difference in means is essentially a visual representation of the cross-validated R^2 statistic of 0.54 mentioned above.

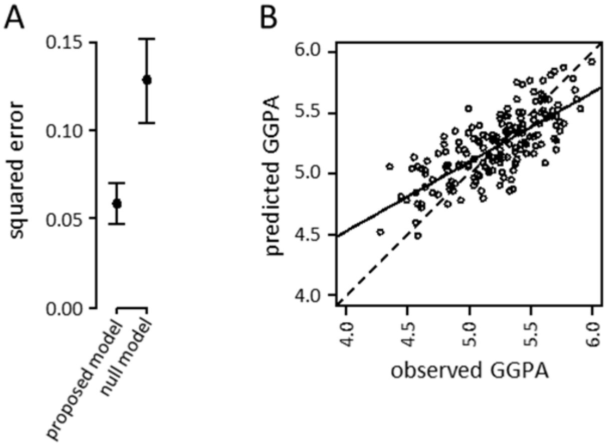

Figure 4: (A) Prediction accuracy (mean and 95% confidence interval) of the model-selection framework and the null model. (B) Observed vs. predicted GGPA. Observed GGPA of individual students is plotted against GGPA as predicted by the model-selection framework. The solid line represents the regression line while the dashed one indicates the 1:1 line.

Figure 4b depicts the observed GGPA of individual students against the GGPA as predicted by the model-selection framework, and, thus, the accuracy. That framework slightly underestimates the range of the observed GGPAs, as becomes apparent if one compares the regression line (solid) with the 1:1 line (dashed).

4.2.2 Prediction performance of individual models and analysis of model instances

To estimate the prediction performance of the eight models, we calculated the cross-validated R^2 statistics for each model (Table 4). For comparisons with traditional approaches, we also trained each model on the entire data and analyzed the resulting model instances for the number of explanatory variables selected and goodness of fit. We used adjusted R^2 statistics in combination with the variance inflation factor. A strong negative correlation existed between the adjusted statistics and the cross-validated R^2 statistics ($r = -0.95$). Indeed, the adjusted statistics sometimes seemed rather off-the-mark, such as in the second row of the table. The variance inflation factor helped us dismissing bad models, but a respective cut-off value needed first to be, somewhat arbitrarily, defined; typical values are either '5' or '10'. Moreover, the combination of adjusted R^2 statistics and the variance inflation factor made a direct comparison of the models' prediction performances difficult (see, for example, PC1 and PC3 in Table 4). It is our understanding that the observed discrepancy is a result of the data-independent penalty introduced by the adjusted R^2 statistics. For this reason, we prefer the cross-validated R^2 statistics.

Table 4: Performance measures and sanity coefficients of the eight individual models. Except for adaptive lasso, all models combine a feature-selection approach with a linear regression model (lm). Values are obtained by analyzing the model instances trained on the entire dataset. Abbreviation: cv-R^2: cross-validated R^2; adj. R^2: adjusted R^2; # sel. var.: number of selected variables; VIF: variance inflation factor.

Model	**cv-R^2**	**adj. R^2**	**# sel. var.**	**VIF**
Forward AIC & lm	0.39	0.62	10	14.0
Backward AIC & lm	0.13	0.66	38	1382.4
Forward BIC & lm	0.51	0.58	3	1.5
Backward BIC & lm	0.37	0.59	7	23.1
Adaptive lasso	0.52	-	3	-
PC1: partial correlation & lm	0.53	0 .54	1	1.0
PC2: partial correlation & lm	0.54	0.57	2	1.3
PC3: partial correlation & lm	0.51	0.58	3	1.5

The cross-validated R^2 statistics in Table 4 suggest that model PC2, which selects two explanatory variables using partial-correlation coefficients, performs best. To assess the statistical significance of its superiority and estimate the uncertainties in prediction performance, we computed the squared errors of all models' GGPA predictions (Figure 5a). We then applied *t*-tests on the distributions of the squared errors. Although PC2 performance was statistically indistinguishable from that of PC1 and adaptive lasso, it significantly outperformed all other models ($p < 0.05$; pairwise *t*-tests, Bonferroni-corrected for multiple testing). This demonstrates that choosing partial correlations for variable selection is appropriate for these data.

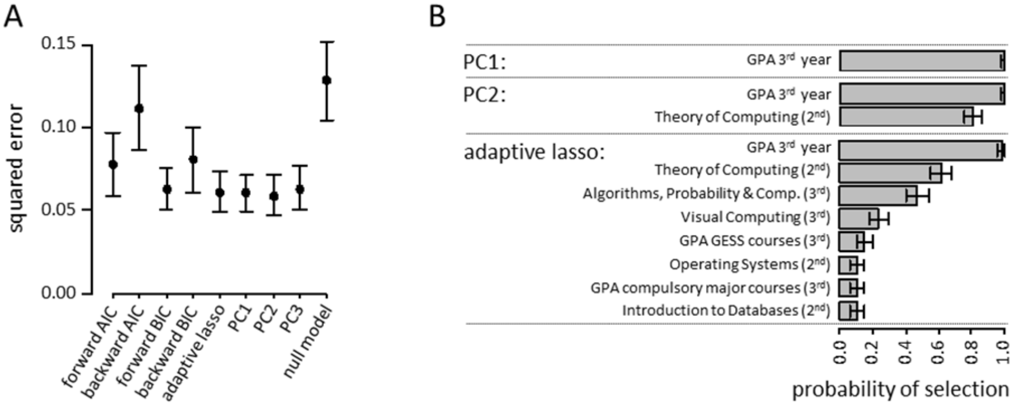

Figure 5: (A) Prediction accuracy of individual models. Means and 95% confidence intervals of squared errors for different models and the null model are presented. (B) Importance of explanatory variables. Probability of selecting an explanatory variable for the three best models is shown with error bars. Explanatory variables are included if they have selection probabilities of at least 0.1.

The model instances of these three best performing models, when trained on the entire set of data, were then investigated for the variables selected and their individual contributions. This was achieved by analyzing the standardized β-coefficients and applying an ANOVA (Table 5). All three models selected the third-year GPA and assigned it by far the greatest weight. Model PC2 and adaptive lasso both chose also *Theory of Computing*, a second-year course, while adaptive lasso identified a third variable, the third-year course *Algorithms, Probability and Computing*. Whereas the contribution of the first two variables was significant in all models, we did not observe any significant contribution from that third variable and so considered it negligible. Moreover, it suffered from 65% missing values. Thus, we treated the selection of this variable with caution since the extent to which it was biased due to values missing and not at random was not clear. Because all models showing statistically indistinguishable performance were quite similar, we were less concerned about type II errors, which would have prevented us from rejecting the null hypothesis of equality because of too-low test power. Since PC1 was the least complex model and exhibited performance results that were statistically indistinguishable from one of the best performing models, we considered it preferable to all others.

Finally, we quantified the stability of variable selection for the three models. For this, we used 200 bootstrap samples and calculated the probability of selection for each explanatory variable and model (Figure 5b). Variable selection proved extremely stable when choosing the first explanatory variable (third-year GPA), quite stable for the second one (*Theory of Computing*), and somewhat stable for the third (*Algorithms, Probability and Computing*). This analysis provided good evidence that, despite the small sample size, we could identify a model that generalized well to new students in the program.

In our analysis, the third-year GPA was by far the most important explanatory variable for predicting GGPA. We found it interesting that, out of all possible GPAs from the third study year – GPAs across course categories and the GPA across the entire study year – the full aggregation was chosen even though it also contained grades achieved in courses not related to Computer Science, e.g., those within the Humanities, or in Social or Political Sciences. The observed dominance of the third-year GPA might have been due to several factors.

Table 5: Variables selected, standardized β-coefficients, and details from ANOVA. Abbreviations: β-coef: β-coefficients; DF: Degrees of freedom; SS: Sum of squares; Significance codes: * p<.05; ** p<.01; *** p<.001

Model Source	**β-coef**	**DF**	**SS**	**F-value**	**Sig.**
PC1					
Third-year GPA	0.76	1	12.02	203.80	***
Residuals		169	9.96		
PC2					
Third-year GPA	0.66	1	12.02	216.53	***
Theory of Computing (second year)	0.12	1	0.64	11.56	***
Residuals		168	9.32		
Adaptive lasso					
Third-year GPA, all performances	0.6	1	12.02	218.13	***
Theory of Computing (second year)	0.09	1	0.64	11.64	***
Algorithms, Probability & Comp. (third year)	0.05	1	0.12	2.24	
Residuals		167	9.20		

The average course in the Bachelor's program has higher selectivity than the average one in the Master's program. Some courses, in particular those at the beginning of the Bachelor's program, put considerable amount of pressure on students and are compulsory. In fact, each year approximately 50% of all students fail the first-year examination. That test may be repeated once; failing twice leads to expulsion. Only in the third year is the selectivity of courses and pressure comparable to those in the Master's program. In addition, students for the first time can choose courses. Typically, they select topics that will prepare them for a major in their Master's program. Furthermore, students' abilities and their adaptation to the academic culture may develop at different rates. Whereas the Master's program is taught in English, the language only gradually switches from German to English in the Bachelor's program. However, the impact of language does not seem to be very strong, because *Theory of Computing*, which is taught in German, has been repeatedly selected. For these reasons, students' performances in the third year of the Bachelor's program faithfully reflect their knowledge and potential for the Master's program.

The other two selected courses – both in the field of Theoretical Computer Science – have an above average degree of mathematical rigor and formalism. Therefore, they are potentially better for quantifying performance when compared with other topics. The course *Theory of Computing* is also taught in a highly standardized manner, always by the same lecturer, and is supported by a self-study book. Moreover, this course offers students two ways to pass the course: successfully completing either two midterm examinations or one final examination, which exerts less pressure on students than typical courses in the second study year. For these reasons, we consider *Theory of Computing* to be rather weakly influenced by construct-irrelevant factors and, thus, a good estimate of a student's capability. This may explain the high selection probability observed here. However, as mentioned previously, considering the third-year GPA alone provides statistically equally good GGPA predictions and is a rather generalizable result.

4.3 Latent structure of the undergraduate program and aggregation of explanatory variables

4.3.1 Specifically aggregated explanatory variables

The following investigation describes how we meet the challenge of improving prediction performance through specific variable aggregations (c.f., Section 3.3.1). In Figure 6a, SqEs are plotted for each set of explanatory variables. The lowest prediction errors were obtained from models trained on year-specific grade averages, or YW. These models significantly outperformed all others ($p < 0.01$; pairwise t-test, Bonferroni-corrected) and, thus, yielded the highest accuracy, thereby supporting our initial hypothesis that partial aggregation presents the most suitable level of detail.

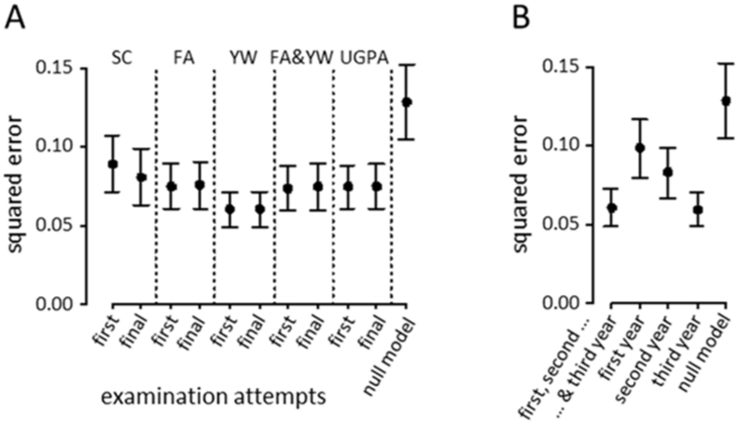

Figure 6: (A) Prediction accuracy of models with different levels of course aggregation. Means and 95% confidence intervals of squared errors are shown for linear regression models trained on various sets of explanatory variables. SC, single course; FA, factor analysis; YW, year-wise clustering; UGPA, undergraduate grade point average. Models were trained on explanatory variables containing either first or final examination attempts. The null model is plotted for comparison. (B) Predictive value of individual Bachelor's program study years. Means and 95% confidence intervals of squared errors are shown for linear regression models trained on GPAs from final attempts in year-wise clustered courses. Sets of explanatory variables contain GPAs from all three study years and individual years.

To assess the contribution of individual study years, we trained linear regression models on year-specific GPAs; three received the single GPA from each year and one received all three. Because prediction performance did not differ between models based on information from first examination attempts versus those based on information from final attempts (p-values between 0.51 and 0.93; pairwise t-tests, Bonferroni-corrected), Figure 6b shows the accuracies on the basis of grade averages from final examination attempts. We recognized that the comparable prediction performance of models trained on first examination attempts and those trained on final examination attempts arose from the fact that the variables only differ from 13% in the first year to 3% in the third year (see Table 3), a situation that did not seem to have sufficient influence on the indicative value.

Of the three models that focused on a single GPA, the one based on the third study year was most predictive for future graduate-level performance. The inclusion of first- and second-year GPAs did not lead to any improvement ($R^2_{\text{change}} = 0.01$, $F_{\text{change}} = 2.7$, $p = 0.06$). Comparing the model based on the third-year GPA with the one based on all three GPAs (first and fourth model in Figure 6b) revealed an insignificant difference ($p = 0.99$; pairwise *t*-test, Bonferroni-corrected). Thus, the indicative value of the third-year GPA could not be improved by adding first-year and second-year GPAs. Importantly, this provided further evidence for the dominance of the third-year GPA, upon which we previously elaborated.

This finding contradicted our expectation because we were surprised to learn that YW aggregation outperformed the FA-based aggregation approach. The Bachelor's and Master's programs in Computer Science at ETH Zurich are most appropriately placed between Mathematics and Electrical Engineering. The content ranges from highly formalized and mathematically oriented theoretical topics such as Algorithmics, Theory of Computing, and Cryptography, to more engineering-oriented and less formal topics such as Programming, Databases, Networks, and Software Engineering. Therefore, we expected to identify a structure that reflects these two domains, a theoretical and a more applied one. This lack of an expected finding motivated us to conduct the following sophisticated investigation.

4.3.2 Latent structure of the undergraduate program

We reasoned that the inferior results from the approach using FA might be explained by a structuring of the undergraduate program that does not reflect a set of independent constructs but primarily assesses a single set of abilities. To test this hypothesis, we ran two post-hoc analyses of the undergraduate program structure.

First, we computed the SVD of undergraduate grades from the first and second years, selecting model-order by following the minimum transfer-cost principle. The third year was not included because the large amount of missing values in the data jeopardized the validity of the results. The model with just one factor generalized best, as shown in Figure 7a, where the model with the lowest transfer costs exhibits an order of '1'. Figure 7b shows the V-matrix of the full-rank SVD of the entire dataset. We observed that, whereas all courses loaded quite uniformly on the first factor (represented by the first column of the V-matrix), the remaining entries seemed to be distributed randomly. Both outcomes support the notion that the first two years of the undergraduate program can best be described as a one-dimensional construct and, thus, assess a unique set of abilities. We assume that this observation holds for the entire undergraduate program as third-year students deepen their knowledge in areas introduced during the second year.

Importantly, in factorization, neither the numbers of factors to be considered nor determining the numbers of factors that generalize best is a straightforward model-selection problem. The minimum transfer-cost principle provides a solution to this problem that is easy to implement and to interpret. It greatly helped us understand the results obtained when using FA.

We also computed Cronbach's α to characterize the consistency with which the undergraduate program assessed the above set of abilities. A value of $\alpha = 0.98$ was obtained for final examination attempts, indicating *excellent* consistency (George and Mallery, 2011). This result provides further evidence that calculating GPAs across any group of courses does not lead to a notable loss of information but instead increases the stability of inference. This is because noise introduced by construct-irrelevant factors is reduced through averaging. However, as demonstrated before, when attempting to predict graduate-level performance, it is beneficial to consider temporal proximity.

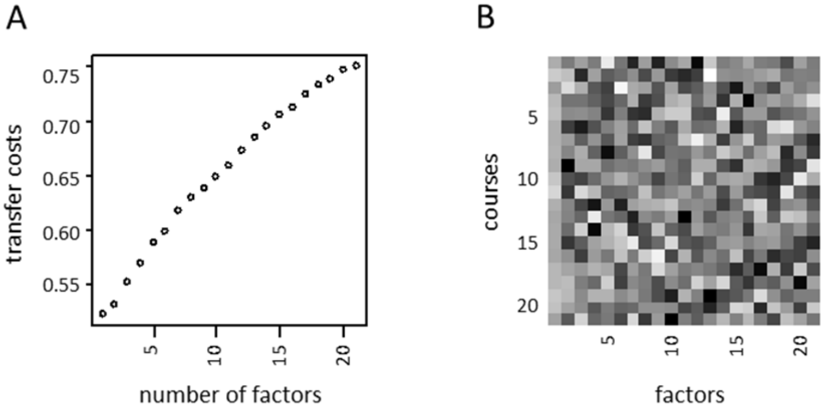

Figure 7: (A) Transfer costs for SVD. Data indicate final examination attempts during the first and second undergraduate years vs. the number of factors considered. (B) Heat map of V-matrix. Loadings of courses on factors are shown, i.e., V-matrix of SVD decomposition, with colors encoding values between –1 (black) and +1 (white).

5. Conclusions

In this paper, we analyzed how well indicators of undergraduate achievements can predict graduate-level performance. We used data comprising 171 student records acquired from the Bachelor's and Master's programs in Computer Science at ETH Zurich, Switzerland. Notably, this dataset was complete, making it possible to render an in-depth analysis, which is generally not feasible when reviewing data for selectively admitted students. We chose linear regression models in combination with different variable-selection methods i) to examine the predictive power of undergraduate-level performance indicators and ii) to explore whether purposeful aggregation of grades further improves prediction performance and understanding. For our analysis, we employed analytical techniques that otherwise have not been widely adopted for educational research.

Our first major result is determining a correlation of 0.65 between the GPA at the undergraduate level and the one at the graduate level. This outcome emphasizes the relevance of indicators of undergraduate achievements for graduate admissions decision-making. When using a model-based approach and considering not only the UGPA but also annual GPAs and single grades, the predictive power increases and indicators of undergraduate performance can explain as much as 54% of the variance in subsequent graduate-level performance. This value represents a notable gain over previous reports of only 4% to 17% explained variance. We attribute this improvement primarily to the completeness of our data, to the strong consecutive nature of the Computer Science curriculum, and to the fact that data were collected within one institution. Therefore, we deduce that this 54% explained variance is probably an optimistic estimate of the upper bound for the predictive value of undergraduate achievements. Within the context of admitting international students, this value is expected to be initially lower than what is reported. However, over time and as experience is gained, it might be possible to control for factors such as differences in language abilities, academic cultures, or curricula. In doing so, we will be able to improve the predictive value of undergraduate performance indicators.

Our second result is that the third-year GPA is repeatedly identified as the most significant explanatory variable. Notably, only the full aggregation was selected from all third-year GPAs, including grades earned in courses unrelated to Computer Science. While this finding is in contrast to the popular view that only subject-related courses are relevant indicators of future graduate-level success, it is consistent with reports from research on the transition from high school to undergraduate studies in Germany (Baron-Boldt et al., 1988; Trapmann et al., 2007). While further evaluation is needed about the generalizability of our result, we suggest that institutions consider using it for programs of comparable consecutive nature, with similar stringent undergraduate degree requirements, and when mainly teaching components are associated with a Master's program.

Our third result is that the partial aggregation approaches based on year-wise clustering of individual undergraduate achievements provide the best predictions. Again, the third-year GPA yields the most accurate predictions, and they cannot be improved by adding first- and second-year GPAs. This result distinctly contrasts with the popular view, also shared by professors, that the most important indicators of excellence are the grades earned in challenging first-year courses in Mathematics. Our results, however, suggest that high selectivity of a course is not necessarily related to its predictive value of future performance, at least for those students who cannot opt out of challenging first-year courses in Mathematics and still complete the program. Furthermore, using FA, one might expect to distinguish between an ability related to Mathematics and one related to Engineering. That these abilities are distinctive is another view often expressed. However, we observed that the latent structure of the undergraduate program can best be described as a one-dimensional construct that assesses a unique set of abilities with remarkably high consistency. This finding implies that all courses require about the same set of skills from students. Arguably, the effect of evaluating different capabilities might be minimal in comparison to the confounding factors introduced, for example, by examiners, or else the sample size might be too small to detect a more complex structure.

Proper data mining techniques are essential if one aims at predicting graduate level performance in a small sample size setting. We estimated prediction performance employing two layers of cross-validation, which prevents information-leakage from the training and validation phases to the testing phase. Furthermore, we used adaptive lasso, which was proven to perform competitively in high-dimensional data settings. It matched the performance of the approach that relied upon partial correlations, which we preferred for its simplicity. Bootstrapping is highly valuable to identify significant explanatory variables, especially when they appear in different bootstrap samples and different models. In summary, our approach provides considerable reassurance that undergraduate achievements are highly indicative of graduate-level success (54% explained variance) and that the third-year GPA is the most important explanatory variable.

When relatively recent methodological approaches were compared to more traditional ones for estimating explained variance while coupling variable selection and linear regression, the adjusted R^2 statistics seemed to return inconsistent results possibly because of the data-independent penalty introduced by those statistics. While the variance inflation factor would lead to dismissing the worst models, its use is still questionable in the set of acceptable models. Thus, to enhance the credibility of their interpretations, we strongly suggest that both researchers and practitioners rely instead on cross-validated R^2 statistics and use bootstrapping to test the stability of their results. While not crucial to the development of this paper, we find that adaptive lasso is one of the best models, outperforming traditional approaches that utilize AIC or BIC for model selection with few samples per parameter. Its simultaneous variable selection and parameter estimations, as well as its oracle property, make adaptive lasso an attractive candidate. Finally, we can also report that optimizing the level of aggregating variables is particularly useful when high multicollinearity is present in the data.

APPLICATIONS AND FUTURE DIRECTIONS. Our results demonstrate that admissions committees can rely on undergraduate-level performance indicators as selection instruments for Master’s programs in Computer Science within a comparable context. However, our findings also show that, even in this rather ideal setting, where a complete dataset has been collected within a single institution, additional admissions tools are required. Furthermore, we recommend that committees do not rely on grades from single courses but instead look at partially aggregated undergraduate grades. From experience, we know that this is not an uncommon issue of discussion during committee meetings.

As reported, the third-year GPA outperforms the frequently used UGPA by 27%, with the latter variable explaining no more than 42% of the variation in the GGPA. Whether the third-year GPA is the most important indicator at other institutions or within the context of international student recruitment remains to be seen. However, its strong predictive performance in combination with the simplicity of the model makes it an attractive candidate for future analyses. Where such generalizability is demonstrated, our findings support the derivation of admissions rules, depending on the committee’s main objective. For instance, if one wishes to select those graduate-program applicants who are expected to perform above a certain level then one must determine the required third-year GPA based on available data and establish a respective threshold. When an admissions policy is restricted to a fixed number of students, one must rank applicants according to their third-year GPAs and admit the top candidates. Our future research will explore how these current results can be generalized within the context of international student recruitment and how standardized tests such as the GRE® might lend additional support.

6. References

AGBONLAHO, R. O., AND OFFOR, U. J. 2008. Predicting success in a Master of Information Science degree programme. *Education for Information* 26, 3-4, 169-190.

AKAIKE, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 9, 6, 716-723.

ASTIN, A. W. 1993. *What Matters in College?: Four Critical Years Revisited*. Jossey-Bass Publishers, San Francisco, CA, USA.

ATKINSON, R. C., AND GEISER, S. 2009. Reflections on a century of college admissions tests. *Educational Researcher* 38, 9, 665-676.

BACKHAUS, K., ERICHSON, B., PLINKE, W., AND WEIBER, R. 2006. *Multivariate Analysemethoden* (11th ed.). Springer, Berlin/Heidelberg, Germany.

BAIRD, J.-A. 2011. Why do people appeal Higher Education grades and what can it tell us about the meaning of standards?. *Assessment in Education: Principles, Policy & Practice* 18, 1, 1-4.

BAKER, R. S. J. D., AND YACEF, K. 2009. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining* 1, 1, 3-17.

BAKER, R. S. J. D., GOWDA, S. M., AND CORBETT, A. T. 2011. Automatically detecting a student’s preparation for future learning: Help use is key. *Proceedings of the 4th International Conference on Educational Data Mining*, 179-188.

BARON-BOLDT, J., SCHULER, H., AND FUNKE, U. 1988. Prädiktive Validität von Schulabschlussnoten: Eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie* 2, 79-90.

BERGIN S., AND REILLY, R. 2006. Predicting introductory programming performance: A multi-institutional multivariate study. *Computer Science Education* 16, 4, 303-323.

BIRKEL, P. 1978. *Mündliche Prüfungen*. Kamp, Bochum, Germany.

BOWERS, A. J. 2011. What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation* 17, 3, 141-159.

BREIMAN, L. 2001. Random forests. *Machine Learning* 45, 1, 5-32.

BREIMAN, L., AND SPECTOR, P. 1992. Submodel selection and evaluation in regression. The X-Random case. *International Statistical Review* 60, 3, 291-319.

BRIDGEMAN, B., BURTON, N., AND CLINE, F. 2009. A note on presenting what predictive validity numbers mean. *Applied Measurement in Education* 22, 2, 109-119.

BÜHLMANN, P. AND VAN DE GEER, S. 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Berlin/Heidelberg, Germany.

CAMARA, W. J. 2005. Broadening criteria of college success and the impact of cognitive redictors. In *Choosing Students: Higher Education Admissions Tools for the 21st Century*, W. J. Camara and E. W. Kimmel (Eds.). Lawrence Erlbaum Associates, Mahwah, NJ, USA, 53-79.

CHO, Y., AND BRIDGEMAN, B. 2012. Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities. *Language Testing* 29, 3, 421-442.

CONLEY, D. T. 2005. *College Knowledge: What it Really Takes for Students to Succeed and What We Can Do to Get Them Ready*. Jossey-Bass, San Francisco, CA, USA.

COSTELLO, A. B., AND OSBORNE, J. W. 2005. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation* 10, 7, 173-178.

CRONBACH, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297-334.

CRONBACH, L. J. 1971. Test validation. In *Educational Measurement* (2nd ed.), R. L. Thorndike (Ed.). American Council on Education, Washington, DC, USA, 443-507.

CUNY, J., AND ASPRAY, W. 2000. Recruitment and retention of women graduate students in computer science and engineering: Results of a workshop organized by the Computing Research Association. *SIGCSE Bulletin* 34, 168-174.

DAWES, R. 1975. Graduate admission variables and future success. *Science* 187, 4178, 721-723.

DE FEYTER, T., CAERS, R., VIGNA, C., AND BERINGS, D. 2012. Unraveling the impact of the big five personality traits on academic performance: The moderating and mediating effects of self-efficacy and academic motivation. *Learning and Individual Differences* 22, 4, 439-448.

DISTEFANO, C. 2009. Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation* 14, 20, 1-11.

DOWNEY, M., COLLINS, M., AND BROWNING, W. 2002. Predictors of success in dental hygiene education: A six-year review. *Journal of Dental Education* 66, 11, 1269-1273.

EACEA. 2012. *The European Higher Education Area in 2012: Bologna Process Implementation Report*. Education Audiovisual & Culture Executive Agency (EACEA), Belgium.

EFRON, B., AND TIBSHIRANI, R. J. 1994. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. Chapman and Hall, London, UK.

EVANS, P., AND WEN, F. K. 2007. Does the medical college admission test predict global academic performance in osteopathic medical school? *Journal of the American Osteopathic Association* 107, 4, 157-162.

FABRIGAR, L. R., WEGENER, D. T., MACCALLUM, R. C., AND STRAHAN, E. J. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods* 4, 272-299.

FETTER, J. H. 1997. *Questions and Admissions: Reflections on 100,000 Admissions Decisions at Stanford*. Stanford University Press, Stanford, CA, USA.

FRANK, M., CHEHREGHANI, M., AND BUHMANN, J. M. 2011. The minimum transfer cost principle for model-order selection. *Machine Learning and Knowledge Discovery in Databases*. Lecture Notes in Computer Science Series 691. Springer, Berlin/Heidelberg, Germany, 423-438.

FREY, K., AND FREY-EILING, A. 2009. *Ausgewählte Methoden der Didaktik*. vdf UTB, Stuttgart, Germany.

GEORGE, D., AND MALLERY, P. 2011. *SPSS for Windows Step by Step: A Simple Guide and Reference*, 18.0 Update (11th ed.). Allyn & Bacon / Pearson, Boston, MA, USA.

GUYON, I., AND ELISSEEFF, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3, 272-299.

GRAHAM, J.G. 1987. English language proficiency and the prediction of academic success. *TESOL Quarterly* 21, 3, 505-521.

HARTNETT, R.T., AND WILLINGHAM, W.W. 1980. The criterion problem: What measure of success in graduate education? *Applied Psychological Measurement* 4, 281-291.

HERZOG S. 2006. Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression. *New Directions for Institutional Research* 131, 17-33.

KAISER, H. F. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika,* 23, 187-200.

KANE, M. T. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement* 50, 1, 1-37.

KEHM, B. M. 2010. The future of the Bologna Process – The Bologna Process of the future. *European Journal of Education* 45, 4, 529-534.

KLAPP LEKHOLM, A., AND CLIFFORDSON, C. 2008. Discrepancies between school grades and test scores at individual and school level: Effects of gender and family background. *Educational Research and Evaluation* 14, 2, 181-199.

KOYS, D. 2010. GMAT versus alternatives: Predictive validity evidence from Central Europe and the Middle East. *Journal of Education for Business* 85, 3, 180-185.

KUNCEL, N. R., ONES, D. S., AND HEZLETT, S. A. 2001. A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin* 127, 162-181.

LANE, J., LANDE, A., AND COCKERTON, T. 2003. Prediction of postgraduate performance from self-efficacy, class of degree and cognitive ability test scores. *Journal of Hospitality, Leisure, Sport and Tourism Education* 2*,* 1, 113-118.

LANGFELDT, H.-R., AND FINGERHUT, W. 1974. Empirische Ansätze zur Aufklärung des Konstruktes "Schulleistung". In *Leistungsbeurteilung in der Schule*, K. Heller (Ed.). Quelle & Meyer, Heidelberg, Germany.

MEINSHAUSEN, N., AND BÜHLMANN, P. 2010. Stability selection (with discussion). *Journal of the Royal Statistical Society* B 72, 417-473.

MESSICK, S. 1989. Validity. In *Educational Measurement* (3rd ed.), R. L. Linn (Ed.). Macmillan New York, NY, USA, 13-103.

NEWTON, P. E. 2012. Questioning the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives* 10, 1-2, 110-122.

NUGENT, G., SOH, L.-K., SAMAL, A., AND LANG, J. 2006. A placement test for computer science: Design, implementation, and analysis. *Computer Science Education* 16, 1, 19-36.

OSWALD, F. L., SCHMITT, N., KIM, B., RAMSAY, L. J., AND GILLESPIE, M. A. 2001. Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology* 89, 187-207.

OWENS, M. K. 2007. Executive Education: Predicting Student Success in 22 Executive MBA Programs. *GMAC® Report Series*, RR-07-02.

PEÑA-AYALA, A. 2014. Review: Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications* 41, 4, 1432-1462.

POROPAT, A. 2009. A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin* 135, 322-338.

PRECKEL, D., AND FREY, K. 2004. Ein Überblick über Prädiktoren für Studienerfolg. Retrieved from: http://www.ifvf.ethz.ch/news/index.

RA, J., AND RHEE, K.-J. 2014. Efficiency of selecting important variable for longitudinal data. *Psychology* 5, 1, 6-11.

RAFFERTY, A. N., DAVENPORT, J., AND BRUNSKILL, E. 2013. Estimating student knowledge from paired interaction data. *Proceedings of The 6th International Conference on Educational Data Mining (EDM 2013)*.

RAMSEIER, E. 1977. Determinanten des Studienerfolgs: Zusammenfassung der Ergebnisse einer Befragung des schweizerischen Immatrikulationsjahrganges 1965 in einer Pfadanalyse. *Revue Suisse de Sociologie* 3, 3, 57-73.

RAU, W. 2001. Response: To replicate or not to replicate: Is that Schuman's question? *Sociology of Education* 74, 1, 75-77.

RAU, W., AND DURAND, A. 2000. The academic ethic and college grades: Does hard work help students to "Make the Grade"? *Sociology of Education* 73, 1, 19-38.

RINDERMANN, H., AND OUBAID, V. 1999. Auswahl von Studienanfängern durch Universitäten - Kriterien, Verfahren und Prognostizierbarkeit des Studienerfolgs. *Zeitschrift für Differentielle und Diagnostische Psychologie* 20, 3, 172-191.

RITZEN, J. 2010. *A Chance for European Universities or: Avoiding the Looming University Crisis in Europe*. University Press, Amsterdam, The Netherlands.

ROMERO, C., AND VENTURA, S. 2010. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics* 40, 6, 601-618.

ROMERO C., ROMERO J. R., AND VENTURA S. 2014. A survey on pre-processing educational data. In *Educational Data Mining*. Studies in Computational Intelligence 524, A. Peña-Ayala (Ed.). Springer International Publishing, Switzerland, 29-64.

SCHUMAN, H. 2001. Comment: Students' effort and reward in college settings. *Sociology of Education* 74, 1, 73-74.

SCHWARZ, G. 1978. Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461-464.

SCIME, A. 2008. Globalized computing education: Europe and the United States. *Computer Science Education* 18, 1, 43-64.

SOMMERLA, G. 1976. Zur Praxis der Leistungsfeststellung und Bewertung in der Schule. *Westermanns Pädagogische Beiträge* 28, 450-461.

SUELLWOLD, F. 1983. Pädagogische Diagnostik. In *Enzyklopädie der Psychologie*. Themenbereich B: Methodologie und Methoden. Serie II: Psychologische Diagnostik. Band 2: Intelligenz- und Leistungsdiagnostik, L. Michel (Ed.). Hogrefe, Göttingen, Germany, 307-386.

TENT, L. 1969. *Die Auslese von Schülern für weiterführende Schulen: Möglichkeiten und Grenzen*. Beiträge zur Theorie und Praxis. Hogrefe, Göttingen, Germany.

THEOBALD R., AND FREEMAN, S. 2014. Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research. *CBE-Life Sciences Education* 13, 41-48.

THURSTONE, L. L. 1947. *Multiple-factor Analysis*. The University of Chicago Press, Chicago, IL, USA.

TIMER, J. E., AND CLAUSON, M. I. 2010. The use of selective admissions tools to predict students' success in an advanced standing baccalaureate nursing program. *Nurse Education Today* 31, 6, 601-606.

THORSEN, C. 2014. Dimensions of norm-referenced compulsory school grades and their relative importance for the prediction of upper secondary school grades. *Scandinavian Journal of Educational Research* 58, 2, 127-146.

THORSEN, C., AND CLIFFORDSON, C. 2012. Teachers' grade assignment and the predictive validity of criterion-referenced grades. *Educational Research and Evaluation* 18, 2, 153-172.

TRAPMANN, S., HELL, B., WEIGAND, S., AND SCHULER, H. 2007. Die Validität von Schulnoten zur Vorhersage des Studienerfolgs: Eine Metaanalyse [The validity of school grades for predicting academic success: A metanalysis]. *Zeitschrift für Pädagogische Psychologie* 21, 11-27.

TRUELL, A. D., ZHAO, J. J., ALEXANDER, M. W., AND HILL, I. B. 2006. Predicting final student performance in a graduate business program: The MBA. *Delta Pi Epsilon Journal* 488, 3, 144-152.

VENTURA JR., P. R. 2005. Identifying predictors of success for an objects-first CS1. *Computer Science Education* 15, 3, 223-243.

WIKSTRÖM, C., WIKSTRÖM, M., AND LYRÉN, P.-E. 2009. Prediction of study success: Should selection instruments measure cognitive or non-cognitive factors? In *Assessment for a Creative World*. Paper presented at 35th Annual IAEA Conference, Brisbane, Australia, 13-18 September 2009.

WILLINGHAM, W. W. 1974. Predicting Success in Graduate Education. *Science* 183, 4122, 273-278.

WILLINGHAM, W. W., YOUNG, J. W., AND MORRIS, M. M. 1985. Success in college: The role of personal qualities and academic ability. College Board Publications, NY, USA.

ZOU, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101, 476, 1418-1429.