Addendum

Summary and Valorization
Summary and Valorization

“At this point in our history, [...] we can process exabytes of data at lightning speed, we have the potential to make bad decisions far more quickly, efficiently, and with far greater impact than we did in the past.”

Susan Etlinger. What do we do with all this big data? TED@IBM
(San Francisco, 2014)

As stated in the introduction, the aim of this thesis is to understand nuisances and specific features in education data better, as a means to improve empirical research and evidence-based policy making. In the above mentioned quote, Susan Etlinger stresses the importance of being careful not to decouple data generation from data analysis, and therewith, from evidence-based decision-making. This is all the more important, since more and more data become available. However, data out of context lose their value. This holds generally as well as in education.

This thesis investigates the advantages as well as the pitfalls of the broad availability of test scores over the course of a school career and of combining data from different sources. The economics of education research group at Maastricht University and the OnderwijsMonitor Limburg provided the perfect conditions for this purpose. The researchers involved in the design and the realization of the panel study are members of the research group. In the course of writing this thesis, they could answer also detailed questions regarding the local context as well as the process of data collection; some of them were involved as co-authors for a specific study. Strong relationships with the participating schools as well as policy institutions are maintained until today. Additionally, researchers of different disciplines were involved and provided diverse perspectives on the topics. All this allowed to carefully consider and analyze in how far more or less detailed knowledge of the context or more or less extensive series of test scores matter for the results reached. This is what is meant by the sub-title “Test Scores and Beyond”.

This addendum provides a summary of the thesis and the implications arising from the results. The overarching research question is in how far we can trust empirical research results in the economics of education field to reflect relationships and in the best case causal links that apply to the whole population. In order to get closer to answering this question, the different chapters investigate several potential flaws of data: selective participation in research and testing (Chapter 2), non-response bias (Chapter 3), measurement error (Chapter 4) or the diverse content of a test and how using sum scores of tests differs from item level data (Chapter 5).
Summary and Valorization

At first sight this might be quite technical. At a second thought, knowledge about these potential flaws are essential for the link between research and practice via evidence-based innovations or reforms. In the following, the results of the studies compiled in this thesis are summarized and their specific contribution to link research and practice are discussed. The implications are distinguished by relevance for different groups, especially researchers, policy makers, and schools. To end with the most general implications, the order of the chapters is reversed, starting with Chapter 5 and ending with Chapter 2.

Chapter 5 combines economic and psychological research interests in early numeracy skills and has two research questions. It investigates how important it is to have detailed knowledge about the content of an early numeracy test in kindergarten, when predicting later mathematic achievement. This is particularly important, since economists often only use one score as the outcome of a test and disregard differences between items. In order to test, whether differentiation matters, it is first tested, how many different skills can be distinguished. Using an existing test, which was administered in Dutch kindergartens, provides us with a strong empirical basis for this analysis. It enables us to analyze a large sample of more than 3,000 kindergarten children, and a longitudinal sub-sample of 797 children. As a result, five different skills can be distinguished. This differentiation is more detailed than the ones proposed in earlier studies. Taken together, these five early numeracy skills matter for mathematic achievement 6 years later, explaining about 18 percent of the variance. However, whether one sum score or sub-scores for each factor are used for prediction makes almost no difference.

In Math, compared to reading, a lot less is known about early precursors of later achievement. In that sense, the differentiation of early numeracy skills in this study is basic research. Further research could help identify a connection between certain early numeracy skills and later mathematic difficulties. In a third step, this might allow to design targeted and potentially earlier interventions for students with mathematics difficulties.

Beyond this, the study has some direct implications as well. In the economic approach to investigating educational questions tests outcomes are often measured by one sum score. Whether this score adequately represents the skill of interest is rarely empirically supported. The results of Chapter 5 show that, even though a distinction of different early numeric competencies can be made, it does not improve predictive power for a mathematic achievement test six years later. Thus, the results support the practice of using sum scores in early numeric tests to predict later achievement tests.
Another more general implication stems from the collaborative interdisciplinary approach. The research project, which was the foundation for the data set, was run by an economic research group. The research question and the methodology stem from psychology. Before collaborating, economic researchers did not think about using the data in this way and looking at test scores on item level. Conversely, in this specific field of psychology, researchers had not considered using data of existing tests, rather than conducting their own specific study before. Thus, this study shows how collaboration can open up new opportunities and save time and money by sharing resources.

Chapter 4 focuses on using test score series to draw conclusions on the development of social mobility. Using a series of comparable mathematics test scores over the course of six years of Dutch primary education, we study how the influence of socio-economic background, measured by the highest parental education level, develops, and how this relates to measurement error. We find that the relationship of parental education level and test scores becomes stronger, measured in terms of increasing correlations. Our results suggest that in the first four years the main driver of this stronger relationship is more precise measurement. When correcting for this, the influence of parental education is first constant or even weakens. However, towards the transition to secondary school it steeply increases. The results of the last regular low stakes test depend most strongly on parental education levels. Performance in the final objective and external test is less dependent on socio-economic background. As we show, correlations are especially vulnerable to the sample composition, while our proposed structural equation model produces very robust estimates. Furthermore, we provide evidence for an equalizing influence of school during the first three years. Additionally, we analyze separate parts of the distribution. Thereby, we uncover that the increase of social inequality later on is mostly driven by students performing below average.

The main implication for researchers and policy makers is to be careful with the identification and attribution of changes in social inequality and social mobility. Not every increase in correlations between the socio-economic background and children’s performance relates to an increased dependency. Especially in the first few years of school such an observation is likely to be driven by increased precision in measurement. There are different potential reasons for this: students getting used to the testing situation, ongoing skill development or changes in test length and content, just to name some. Before holding schools responsible for such developments, researchers and policy makers should consider and test these alternative explanations. Otherwise, they might unintentionally punish schools with a disproportionately high fraction of students from a low socio-economic background.
The study has further implications regarding the design of transitions in education systems. Shortly before the transition to secondary school, students take one more low stakes test, which is graded internally, and one high stakes test, which is graded externally and unknown to the teachers. We find that students’ socio-economic background plays a smaller role in determining performance in the latter than in the former test. This supports using objective tests for important transition decisions, rather than tests known, conducted and graded by the teachers. Just recently, Dutch policy makers made the internally graded test more important for the transition decision than the externally graded one. Based on our results, this will likely lead to an increase in social inequality in secondary school placement.

Finally, schools aiming to foster social mobility can use some of the results on the overall development and the within-year pattern of social inequality. Social-inequality seems to evolve in the years before transitions. Thus, this is the time when badly performing students of low socio-economic backgrounds need most support. Additionally, schools could counteract the identified within-year pattern of increasing social inequality over the summer. They could improve social mobility by targeting students from low socio-economic backgrounds directly after the summer holidays or even during the summer holidays, by offering summer programs.

In Chapter 3, we investigate whether and how the fact that not all parents respond to surveys sent out for research purposes influence estimates of intergenerational mobility. Many studies on intergenerational mobility are survey-based and parents’ or schools’ participation in surveys may be non-random. Additionally, only a specific subset of the responding parents fills in the questions of interest. This holds specifically for sensitive data, such as their income. The problem is that, when the characteristic, in which the survey participation is non-random, is correlated with the relationship of interest, this selection cannot be corrected for by weights. However, usually this issue cannot be investigated, because non-responders are not observed. Linking Dutch survey data to administrative income data, we investigate whether selective response behavior biases the estimated relationship between parental income and children’s schooling. We find that the relationship is attenuated because specific schools refuse to participate in the data collection endeavors. Parental survey non-response shifts coefficients in the same direction, but not significantly. Parental item non-response on income questions, on the contrary, biases the estimates upward.

To our knowledge, this is the first study to unveil the effect of non-response on estimates of intergenerational mobility. Thus, it is not clear whether the same results would be found in different contexts. Keeping this in mind, some implications are presented in the following.
Summary and Valorization

Our first contribution to survey-based research in intergenerational mobility is, that we identify three different sources of non-response: parental non-response, school non-response and item non-response. We show that they have to be considered separately, because the effects on the estimates of intergenerational mobility are distinct. Contrary to what is often assumed, researchers do not need to worry too much about individual non-response. It leads to an underestimation of the relationship between parental income and student performance. However, the magnitude of the underestimation is smaller than for institutional non-response and in our sample the bias induced by individual non-response is insignificant. This holds, as long as the studies are based on objective income measures, such as administrative information. If subjective measures of income are used, item non-response will likely lead to a severe overestimation of social mobility.

The non-response rates of published studies on intergenerational mobility are substantial (among the ones cited in Chapter 3 up to 42 percent) and also access to administrative data is no guarantee for a full sample. If biases due to non-response are not taken into account, the results might be misleading. Our results give insights in whether to expect an over- or underestimation and provide an order of magnitude. In conclusion, this study helps to judge empirical evidence on social mobility with respect to parental income. If the parents report their income themselves, the results should be considered to be a lower bound. If an objective measure of parental income is available, the estimated relationship between parental income and student performance is more likely an upper bound.

Chapter 2 provides the introduction to the data set used in all other chapters, the education monitor Limburg, to be exemplary for complex data sets in education. As often in education, it relies on a number of different data sources. Among them are test scores, which are increasingly used to assess and compare student and school performance. As is shown, the completeness of survey and test score data depends crucially on the collaboration of a number of different actors. However, with a specific research question in mind, quality and representativeness of the data often fade into the background. In particular, we examine how representative the participating schools and the covered student population are, as well as how selective schools test and provide test scores.

The main implication is not to take data as granted. The first step should always be understanding the data. We provide an example analysis and a framework to start with. What is the context (region, education system)? How were the data collected? Which different data sources were combined? What is the population of interest? How well is it covered? How
Summary and Valorization

representative is the coverage? What are reasons for missing values – do they not exist, have they not been collected? If the latter is the case, who did not participate? We also cannot not answer all these questions in detail. However, we show that posing them contributes to a better understanding of the data and sometimes even raises new research questions on its own. Every data analysis is only as good as the contextual knowledge it is based upon.

Thus, organizations who conduct surveys should provide extensive contextual information. However, they have an incentive to promote their product as best as possible. Admitting high non-response rates or other flaws, which are important for the interpretation, collides with this. Consequently, those who pay for the surveys (government agencies, universities or in other contexts foundations or corporations) should ask for these details and those who use the data (researchers, policy-makers) should do the same. Researchers, if involved in the design of the survey, can also directly contribute to this.

Another important set of contributors to good data in education are schools. Testing and participating in research and doing this fully increases the validity and thereby the value of that research. However, there is a limit to what schools can do beyond their primary task. Thus, along with this goes an appeal to researchers and policy makers to collaborate in data collection, to bundle their efforts for data quality and to share the resulting data.