Low stakes, high stakes: The predictive power of math achievement tests

Lex Borghans, Leo Kockelkorn, Trudie Schils
Maastricht University School of Business and Economics

March 1, 2013

Abstract
Does what is at stake make a difference for test scores? The impact of personality on test scores depend on the circumstances in which the test is taken, and these influences might differ between low and high stakes tests. This raises the question which personal characteristics are associated with a good performance on either the low or high stakes test and which kind of test is a better predictor of future performance. This paper investigates empirically how the difference between scores from low and high stakes math tests is associated with measures of personality and how these tests differ in predictive power. We use unique data from Dutch primary schools that provide us with information about test scores of pupils on math, including a large set of low stakes tests taken between the age of 4 and 12, a high stakes test taken at the end of primary school (age 12), information about personality traits of these children and outcomes at high school. Our main findings are that motivated and performance-driven students score substantially better on the high stakes test than predicted by the low stake tests. Neurotic children perform less well than expected on the high stakes test than predicted. The low stakes test predicts test scores and school track in secondary school better than the high stakes test. This suggests that the personality traits motivation and neuroticism load more on the high stakes test than would be optimal to predict future performance of the children.

We like to thank Magnus Johannesson, Erik Lindqvist and seminar participants at the Stockholm School of Economics and the ORD in Wageningen for their useful comments.

1 Introduction

There is growing evidence that achievement tests do not only load on skills that they aim to measure such as math or language proficiency, but also on personality traits and motivation. The impact of personality on test scores can depend on the circumstances in which the test is taken. These influences might differ between low and high stakes tests. This raises the question which personal characteristics are associated with a good performance on either the low or high stakes test and which kind of test is a better predictor of future performance.

The aim of this paper is to investigate empirically how the difference between scores from low and high stakes math tests is associated with measures of personality and how these tests differ in predictive power. We use unique data from 120 Dutch primary schools that provide us with information about test scores of pupils on math, including both a large set of low stakes tests taken between the age of 4 and 12, and a high stakes test taken at the end of primary education (age 12). These test data are matched to a student questionnaire at age 12 to analyze the link with personality traits and to data from secondary schools which provides us with outcome measures about later performance of the pupils in secondary education (at age 15).

The analysis is divided into two parts. First, we use the low stakes test to predict the performance of the pupils on the high stakes test using a growth model and analyze the difference between the both measures. Second, we compare the predictive power of the composite low stakes test and the high stakes test. Our main findings are that motivated and performance-driven pupils score substantially better on the high stakes test than predicted by the low stake tests. Neurotic pupils perform less well than expected on the high stakes test than predicted. The low stakes test predicts test scores and track choice reached in secondary school better than the high stakes test. This suggests that the personality traits motivation and neuroticism load more on the high stakes test than would be optimal to predict future performance of the pupils.

A growing body of empirical literature shows that achievement tests load quit substantially on personality traits. A study of Eklöf (2007) on the test-motivation of Swedish students in TIMSS 2003 shows that some students report competitive, comparative or social-responsibility reasons for being highly motivated to perform well on the test, while others are more intrinsically motivated to do so. Duckworth et al. (2011) show that a lack of test motivation is especially observed for students that have below-average IQ. Borghans and Schils (2012) show that the performance on achievement tests substantially drops during the test, and that this performance drop is unrelated to the question difficulty. They also show that

the size of this decline is related to personality traits, mainly agreeableness and motivational attitudes towards learning, which suggests that achievement tests pick up something else than just cognition. Borghans et al. (2008) show that improvements in IQ tests due to monetary incentives differ between students with different personalities. Students with favorable personality traits such as high performance motivation and an internal locus of control perform relatively well in the absence of rewards.

For low and high stakes tests other personality traits might be important. Segal (2011) shows that scores on the AFQT increase when the test is high stakes. Attali et al. (2011) show that males exhibit a larger difference in performance between the low and high stakes tests than females, and that whites exhibit a larger difference in performance between the low and high stakes tests relative to Asians, blacks, and Hispanics. They also find that the larger differential performance between low and high stakes tests among men and whites can be partially explained by the lower level of effort invested by these groups in the low stakes test.

In addition, there is also a discussion in the literature about the effects of testing on education. Some people are worried that high stakes tests will lead to a narrow focus on activities that are tested, both by teachers and students. Using the metaphor of deterring drivers from speeding, Lazear (2006) shows that the optimal rules for high stakes testing depend on the costs of learning and of monitoring. For students that learn at high costs this means announcing the exact requirements of the test. For more able students, though, being more unclear about the exact tested material produces better results. The analysis of Neal and Whitmore Schanzenbach (2007) suggests that the choice of the proficiency standard in accountability systems determines the amount of time that teachers spend on students of different ability levels. Using reforms in the proficiency standard, they find that improvements in math scores seem to be targeted on students in the middle of the distribution, and no or weak gains are found at the upper or lower end of the distribution. Borghans and Schils (2012) develop a theoretical model that shows that the more precise an achievement test gets, the more effort a student puts in the tested activities, and the lower the effort he puts on non-tested activities. The overall result is lower predictions on a student's human capital level, which is a combination of the tested and non-tested activities, at a given test score.

2 Data

The data used for this paper are from the Netherlands. In the Netherlands, primary education starts at the age of 4 (compulsory as of age 5) and ends at age 12 (6^{th} grade). The first two years of primary education (i.e. before 1^{st} grade) were previously separated as kindergarten,

but are now integrated into the primary schools. In these first two years of primary education the pupils learn in a playful manner. The emphasis is on the development of social skills and through puzzles and games they learn to write, count and solving problems. As of age 6 (1st grade), the pupils get formal teaching in reading, writing, and calculating. In the final year of primary education (6th grade), there is three-day standardized achievement test (CITO test) which is very important to determine in which track the pupils are selected for secondary education. This test is not (yet) compulsory for all schools, but about 85 percent of the schools take the test. There are three main tracks in secondary education in which a pupil can be selected:

- Vmbo: 4 years, preparation for lower and medium vocational education, 55 percent of the pupils end up here. This track is further divided into a low (mostly practical education) and a high (mostly theoretical education) track. We call these T1 (low) and T2 (high).
- Havo, or T3: 5 years, preparation for higher vocational education, 25 percent of the pupils end up here; and
- Vwo , or T4: 6 years, preparation for academic education, 20 percent of the pupils end up here.

For the analysis we use a large and unique dataset on the educational development of pupils in a given region of the Netherlands, collected by Maastricht University. These data are collected in a cooperative project with schools, schools boards and municipalities to analyze school performance in order to foster educational improvement. For about 200 primary schools, covering 98 percent of the region, we have information from a questionnaire among children and their parents, test results and IQ test in grade 6, including the high stakes achievement test at age 12 mentioned earlier (CITO test). For about 120 of these school we also have information from the school's Student Tracking System (the current preliminary version is based on one third of these observations). This Student Tracking System includes a wide variety of standardized low stakes tests that the pupils took between the age of 4 to 12, in various domains such as math/calculating, reading and language. To be more specific, we use the following tests from the Student Tracking System in primary education:

- For the first two years, previously known as Kindergarten (aged 4-6), there are two low stakes tests a year (halfway and at the end of the year) on pupils' knowledge of basic calculating concepts, such as larger, more etc.;

- For 1st to 4th grade (aged 6-10) there are two low stakes tests a year (halfway and at the end of the year) on pupils' math skills;
- For 5th grade (aged 10-11) there are three low stakes tests (one halfway and two at the end of the year) on pupils' math skills; and
- For 6th grade (aged 11-12) there is one low stakes test (halfway the year) and one high stake test (near the end of the year) on pupils' math skills.

In addition, at the age of 12 the pupils and their parents also completed a questionnaire with which their personality traits (big5 plus ambition and grit), socio-economic background (SES) and gender is assessed.

The pupils are tested again in 9th grade (around age 15), with a low stakes standardized test. This test is comparable with the PISA test and used to inform schools on their added-value of their student population. The test contains questions from PISA 2006-2009, from COOL 2010-2012 (a Dutch cohort study), and some extra difficult questions from a Belgian match achievement test (provided by the University of Leuven). The test was different for different tracks, with one test for T3 and T4 (the two upper tracks), one for T2 and one for T1. The overlap in questions between the tests enable us, using Item Response Theory, to construct one test score.

This information provides us with multiple observations of similar test domains over the years for the same pupil. However, not for all pupils a complete information set is available. Table 1 shows the number of respondents for the various parts of the data.

Table 1: number of respondents

(1) Low stakes math tests primary education (age 4-12)	N=1,219
(1) plus (2) high stakes math test at the end of primary education (age 12)	N = 532
(1) plus (3) personality traits at the end of primary education (age 12)	N = 701
(1) plus (4) low stakes math achievement test in 9th grade (age 15)	N = 474
(1)-(4), all data available	N = 438

Low stakes math tests during primary school

Figure 1 shows the percentage of pupils that took the several yearly low stakes math tests during primary school. The figure shows that especially in the first year of primary school (age 4-5) participation in math/calculating tests is lowest. There is disagreement among schools in the Netherlands about the usefulness of testing children at this young ages. The

figure also shows a small drop in the participation of tests in between the test ate the beginning of 5th grade (7_1) and the one at the beginning of 6th grade (8_1).

Figure 1: Percentage of pupils that took low stakes math tests in several years

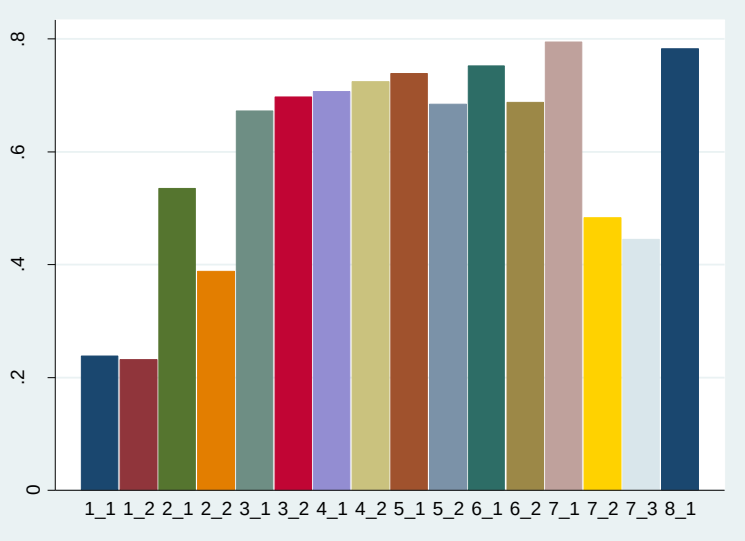
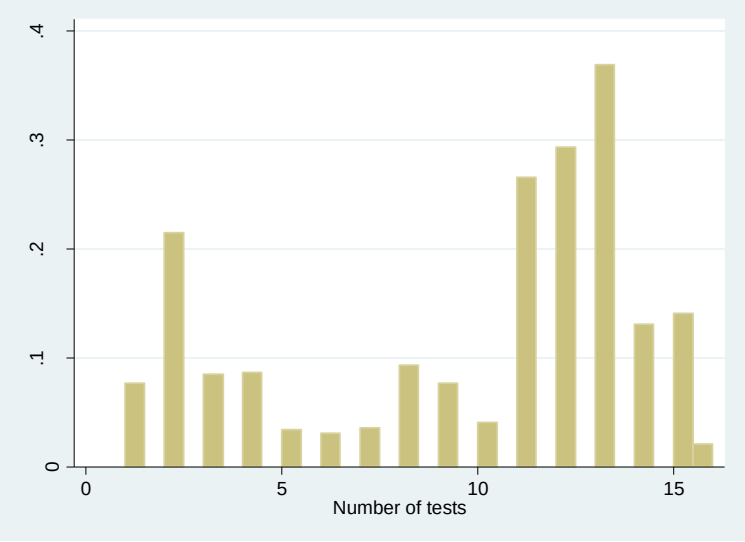


Figure 2 shows the total number of low stakes math tests taken by the pupils. There is quite some variation. Most pupils take 11-13 math tests during the eight years of primary school.

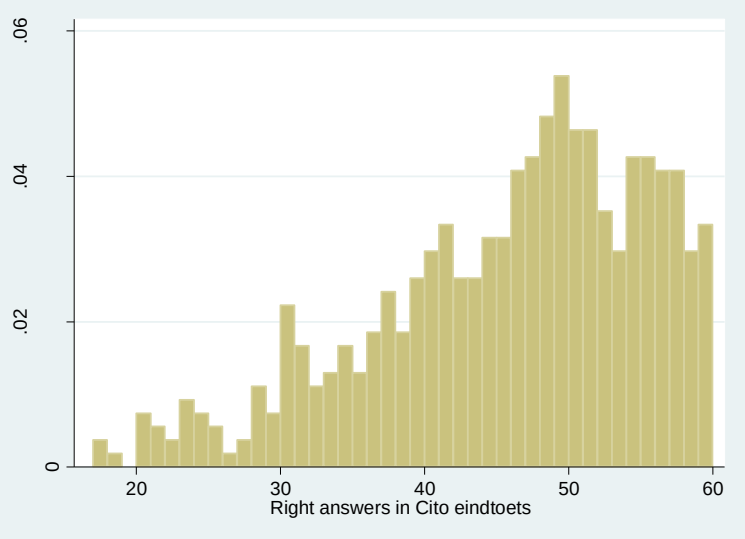
Figure 2: Total number of low stakes math tests taken by pupils in primary schools



High stakes math test at the end of primary school

Figure 3 shows the distribution of the raw test score on the high stakes math test at age 12. The figure shows that there are 60 math questions in the test and that the distribution is somewhat right-skewed (mean = 45.4, median = 47.0 and mode = 49.0).

Figure 3: Distribution of raw test scores on high stakes math test at age 12



Since we do not have complete test information for all pupils, this might raise questions. Could it be possible that the weakest (or best) pupils were not tested as much during their school career than other pupils to avoid demotivation? To get insight in the possible selectivity of the sample, figure 4 shows the distribution of percentile scores for the overall sample and for the subsample of pupils for whom we have complete information. The figure shows that there is no reason to believe that the subsample of pupils for whom we have complete information is a selective sample. The mean score of the overall sample is 58.3 (sd 27.8), and that of the subsample is 58.0 (sd 27.0).

Figure 4: Distribution of percentile scores on high stakes math test at age 12 for two samples

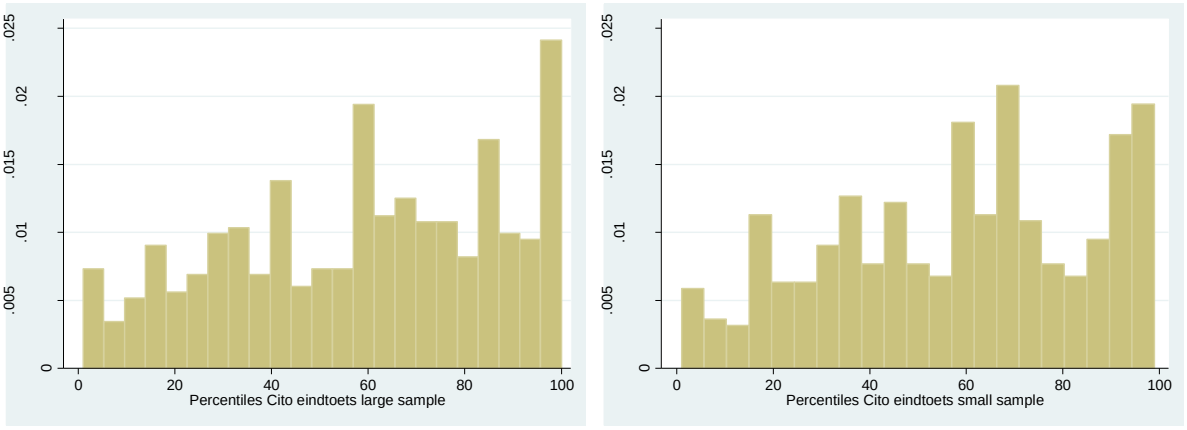
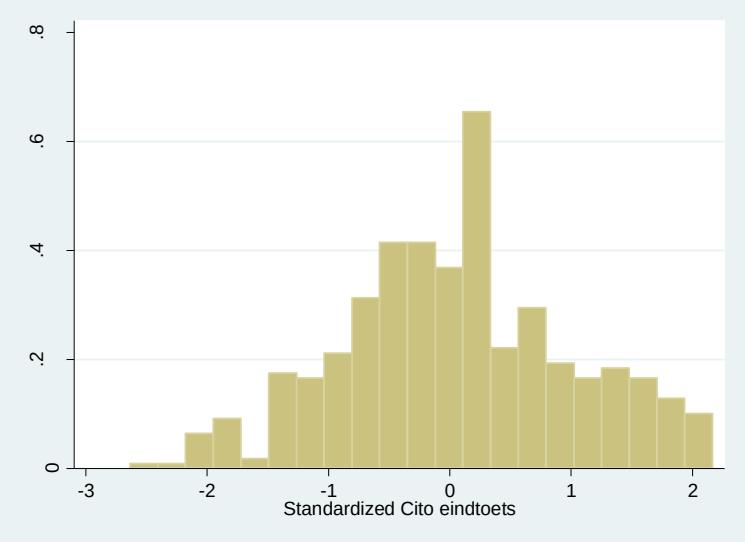


Figure 5 shows the distribution of the standardized test scores for the high stakes test at age 12.

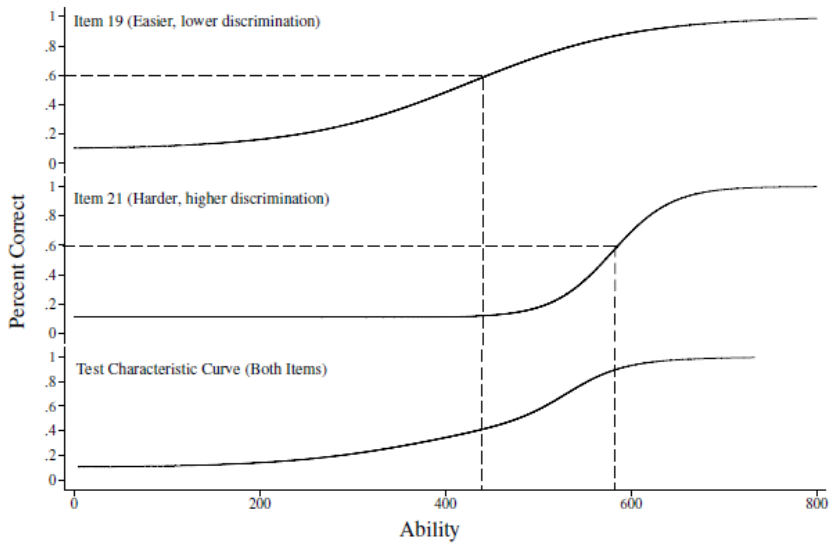
Figure 5: Distribution of standardized test scores on high stakes math test at age 12



Low stakes math test in 9th grade (age 15)

Figure 6 shows the likelihood of a correct answer for two example test questions on the math test in 9th grade, for pupils with different ability levels, using Item Response Theory. For both questions we clearly see that pupils with higher ability levels have a higher likelihood to give a correct answer, and that the curve of the easier test question (item 19) lies above the more difficult question (item 21). The easier question is less useful to discriminate between students as the variation in the likelihood of the correct answer is much smaller.

Figure 6: likelihood of correct answer for two test questions, by pupil's ability level



IRT allows us to use the overlap in test questions between the three different tests (for the different ability levels) to transfer the raw scores into test scores that are aligned on one common scale. Figure 7 shows this transformation. The figure shows that the test scores of the pupils in the highest educational track (T4) are highest and those in the lowest educational track (T1) are generally lowest. Especially between the two middle tracks there is a lot of overlap.

Figure 7: Relation between raw test scores and IRT-scores on math test 9th grade

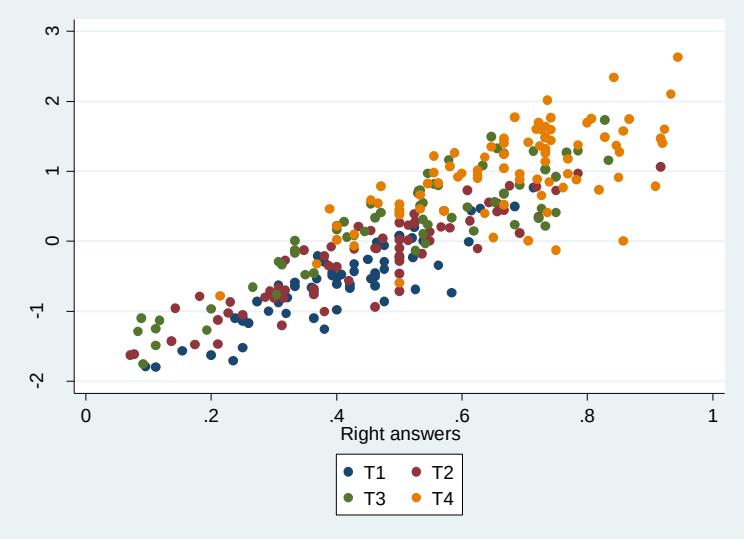
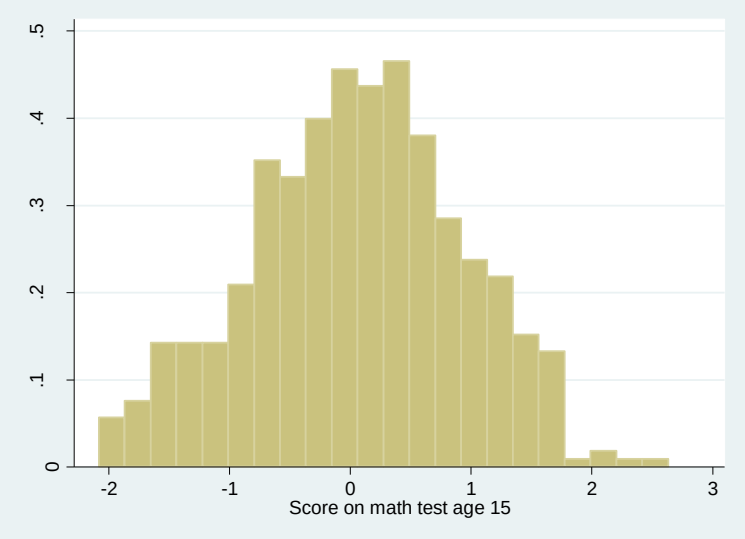


Figure 8 shows the distribution of the IRT-test scores of the math test in 9th grade. There is quite a small right tail, which is due to the fact that we added some extremely difficult math questions in the test to pick up the most excellent pupils.

Figure 8: Distribution of IRT-test scores on high stakes math test in 9th grade



3 Model

We use a structural equation model to analyze the relationship between the (latent) mathematical abilities of a child and the score on the standardized math test. This model is specified as follows:

$$SS_i = \alpha_i + \beta_i L + \epsilon_i$$

where SS_i is the standardized score test in year i, α_i, β_i are the parameters, L is the latent variable representing the child's math abilities and ϵ_i is the test specific error term. In addition, we use an alternative specification, i.e. a growth model, in which we also account for the (latent) growth the pupil develops during the years in his math skills,

$$SS_i = \alpha_i + \beta_i(L + (8 - i)G) + \epsilon_i$$

with G being the latent variable representing the growth of the pupil. In this way we incorporate the possibility that pupils differ in the rate at which they develop their math abilities.

Based on the estimates we predict the score of the high stakes test in 6th grade using the series of low stakes tests and investigate the differences between the several predicted and actual scores and test whether these are related to personality traits.

4 Results

Estimation this Structural Equation Model we obtain estimates for the relationship between the latent ability and the various test scores. Figure 9 plots the parameters of these estimates of the model without the growth parameter for all tests from kindergarten till the last low stakes test in grade 6 and the variance of the error term in these equations. The figure shows that the link between the latent variable and the test score gets stronger over time, while the variance of the error term decreases. Especially the four tests taken at kindergarten are poorer predictors than the later tests. Partially this can be due to the slight difference in content of these tests.

Figure 9 Estimates of the parameters linking the latent variable to the test scores and the variance of the error terms for each low stake test (model without growth)

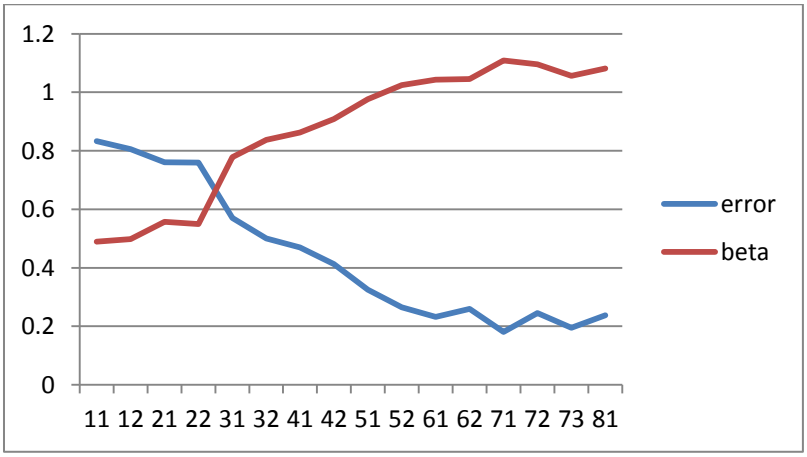


Figure 10 compares the variance of the error term of each equation between the model with and without a latent growth factor. The growth model performs slightly better, especially when the children are young. Figure 11 provides a scatter plot of the predicted latent ability and latent growth, illustrating that the correlation between level and growth is very low.

Figure 10 *Estimates of the variance of the error terms in the model with and without growth for each low stake test*

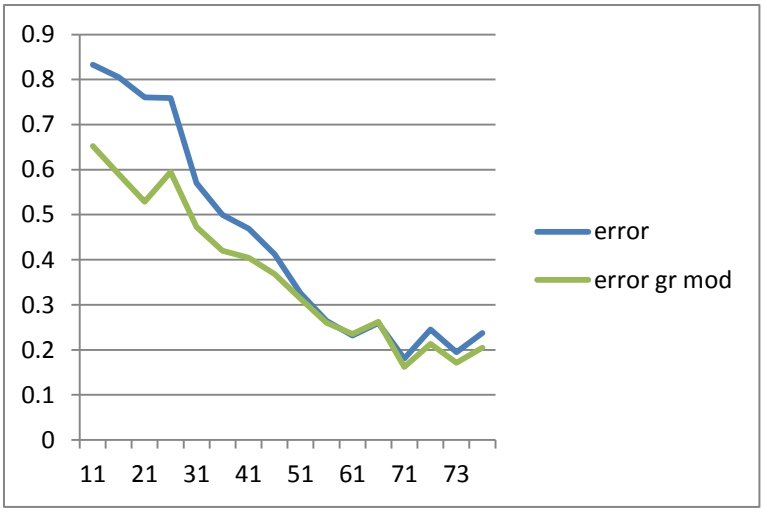
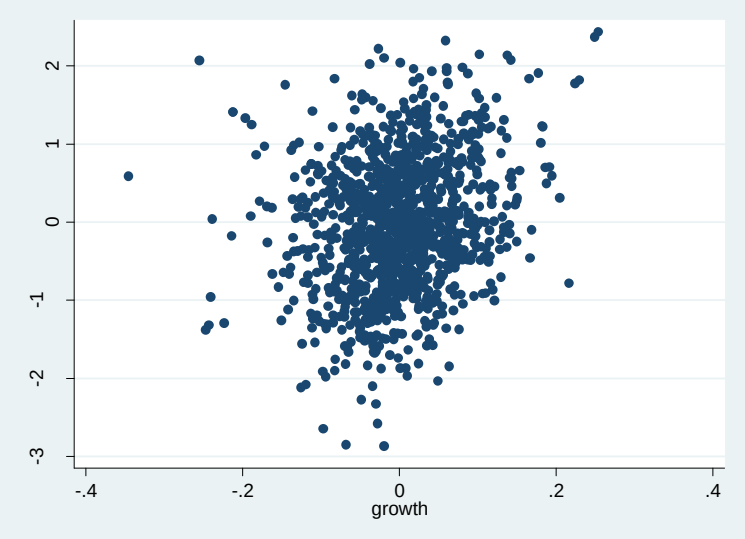


Figure 11 Scatter plot of the predicted level and growth in ability



When a child gets older, gradually more information about his ability becomes available. If more test scores are available the predicted score at the high stakes test improves, while furthermore test takes at higher grade a better predictors. Additionally, the fraction of children that is tested is higher in higher grades. Consequently, the predicted high stakes test score improves when the child gets older. Figure 12 shows the relation between the test score for the high stakes test in 6th grade and the predicted score for the end of the first year of kindergarten year and at the end of 5th grade. The size of the correlation grows over the years, from 0.11 to 0.74. The correlation between the real high stakes test score in 6th grade and the predicted one is 0.90.

Figure 12: Relation between predicted score kindergarten 1 (left) and 5th grade (right) and high stakes test score in 6th grade

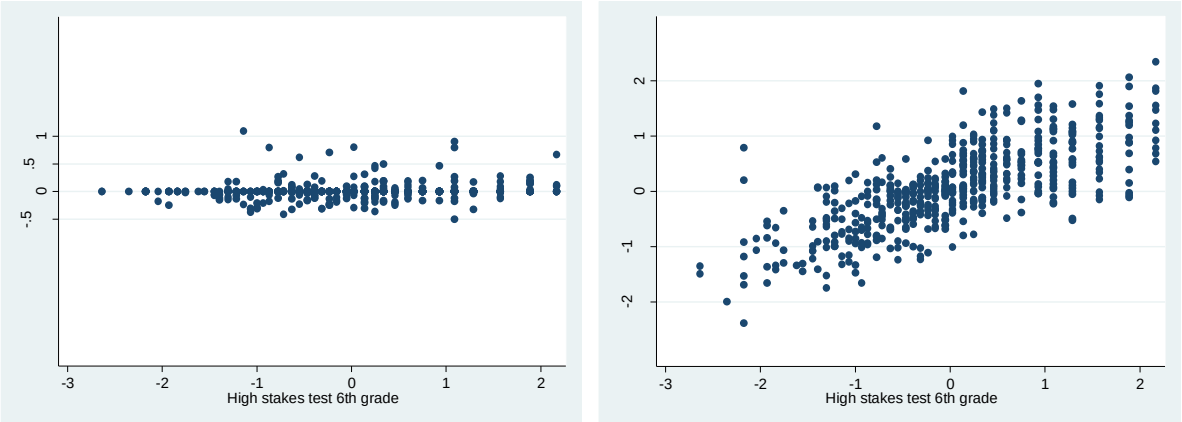


Table 2 shows the estimation results for the relation between the predicted and real test scores and the personality traits. First, the table shows the relation between the personality traits and the real score on the high stakes test in 6th grade. The table shows that conscientiousness and ambition are positively related to the test score and agreeableness is negatively related. Second, the table shows the relation between the personality traits and the predicted score on the high stakes test in 6th grade, using the low stakes test in earlier grades. It shows that there is a small difference in the personality trait that matter: conscientiousness is no longer significant, but openness and grit are. Third, the table shows the relation between the personality traits and the predicted score on the high stakes test in 6th grade, using the growth model of low stakes test in earlier grades. Only ambition now remains correlated with the predicted test score.

Finally, the table shows the relation between personality traits and the difference between the real test score on the high stakes test in 6th grade and the predicted score using the growth model. The table shows that for neurotic pupils the predicted score is higher than the real score, while for ambitious pupils (and only just significant also for more open pupils) the predicted score is lower than the real score. This means that the relation between test scores and personality traits is different when the stakes of the test vary.

Table 2: Estimated coefficients (t-values) for relation between (predicted) test scores in 6th grade and personality traits

	Real score on high stakes test 6th grade	Predicted score on high stakes test 6th grade using basic	Predicted score on high stakes test 6th grade using growth	Difference between [a] and [c]

			model [b]		model [c]			
	[a]							
Openness	0.060	(3.33)	0.063	(5.28)	-0.0014	(-1.32)	0.014	(1.66)
Conscientiousness	0.027	(1.69)	0.004	(0.40)	0.0009	(0.93)	0.009	(1.29)
Extraversion	-0.011	(-0.78)	-0.016	(-1.67)	0.0000	(-0.04)	0.001	(0.16)
Agreeableness	-0.044	(-2.07)	-0.036	(-2.52)	-0.0017	(-1.26)	-0.012	(-1.25)
Neuroticism	-0.009	(-0.72)	0.003	(0.33)	-0.0008	(-1.01)	-0.013	(-2.16)
Ambition	0.035	(3.14)	0.024	(3.16)	0.0019	(2.73)	0.010	(1.97)
Grit	0.017	(1.08)	0.027	(2.59)	0.0007	(0.79)	-0.006	(-0.84)
Constant	-1.569	(-3.04)	-1.345	(-3.83)	-0.0361	(-1.12)	-0.176	(-0.73)

Tables 3 and 4 show the results for the relation between test in 9th grade and the high stakes test in 6th grade as well as the predicted high stakes test in 6th grades using the series of low stakes tests. Both the real test as the predicted test at primary school are very good predictors of school performance at secondary school, as the first two columns of the table show. Combining both the real and the predicted test score in one regression reveals that about 75% of the predicted power loads on the predicted test score rather than the real high stakes test. This suggests that personality factors that affect the high stake test do not improve the predictive power of the test. Predictability would improve if the high stakes test would be less influence by personality factors.

Measurement error might be a concern for this comparison. For that reason we included the link between the test score and the predicted test score with school performance in grade 9 in the structural equation model. This only gradually affects the results suggesting that measurement error does not explain the results in column 3.

Table 3: Estimated coefficients (t-values) for relation between test score in 9th grade and (predicted) test scores in 6th grade

	Using real test score 6th grade (OLS)		Using predicted test score 6th grade (OLS)		Using real plus predicted test score 6th grade (OLS)		Using real plus predicted test score 6th grade (SEM)	
Real test score 6th grade	0.450	(11.36)			0.131	(1.65)	0.125	(1.59)
Predicted			0.573	(13.45)	0.473	(4.61)	0.444	(4.59)

test score 6th grade								
Constant	0.048	(1.17)	0.019	(0.56)	0.032	(0.81)	0.022	(0.66)

Table 4 repeats these analyses for the growth specification. The second and third column of the table differ slightly from Table 3 since the predicted latent variable changes slightly when using a growth specification, but the results remain basically the same. An interesting additional question is whether the estimates growth parameter improves the prediction of later outcomes. This would be very plausible since pupils that follow a steeper growth curve at primary school might maintain this steeper growth at secondary school. We find no evidence though, that this is the case.

Table 4: Estimated coefficients (t-values) for relation between test score in 9th grade and (predicted) test scores in 6th grade

	Using real test score 6th grade		Using predicted test score 6th grade		Using real plus predicted test score 6th grade		Growth	
Real test score 6th grade	0.450	(11.36)			0.073	(0.76)	0.106	(0.10)
Predicted test score 6th grade			0.541	(13.37)	0.508	(4.30)	0.488	(0.12)
Growth							-0.746	(0.61)
Constant	0.048	(1.17)	0.019	(0.56)	0.036	(0.89)	0.034	(0.04)

In several ways we checked the robustness of our results. We used 2PL instead of 3PL to construct our outcome measure, we used the number of right questions instead of normalized score and we used alternative specifications of the growth model, allowing for changes in the growth parameter over time. None of these alternative specification did substantially change the findings.

A limitation of our approach is that the outcome variable in grade 9 is a low stakes math test. Possibly the high stakes test at the end of primary education does not predict better than the score based on low stake tests, because the outcome is not high stake. To look at this, we took the track level that students reached in grade 9 as an alternative outcome. Table 5

provides the results for the model with and without growth. The conclusions are basically unchanged.

Table 5: Track level as outcome variable

	Model without growth		Model with growth	
Real test score 6th grade	0.353	(5.26)	0.244	(3.07)
Predicted test score 6th grade	0.610	(7.14)	0.707	(7.27)
Constant	2.602	(77.29)	2.602	(77.45)

5 Conclusions

In this paper we empirically investigated how the difference between scores from low and high stakes math tests is associated with measures of personality and how these tests differ in predictive power. We use the low stakes test to predict the performance of the pupils on the high stakes test using a growth model and analyze the difference between the both measures. We than compare the predictive power of the composite low stakes test and the high stakes test.

Our main findings are that motivated and performance-driven pupils score substantially better on the high stakes test than predicted by the low stake tests. Neurotic pupils perform less well than expected on the high stakes test than predicted. The low stakes test predicts test scores and track choice in secondary school better than the high stakes test. The growth in math skills during primary school does not predict the performance in secondary school. The evidence suggests that the personality traits motivation and neuroticism load more on the high stakes test than would be optimal to predict future performance of the pupils.

References

Attali, Y., Z. Neeman and A. Schlosser (2011), Rise to the challenge or not give a damn: differential performance in high vs. low stakes tests, IZA DP no. 5693, Bonn: IZA.

Eklof, H. (2007), "Test-taking motivation and mathematics performance in TIMSS 2003." *International Journal of Testing* 7 (3), pp. 311-327.

Borghans, L., H. Meijers and B. Ter Weel (2008), "The role of noncognitive skills in explaining cognitive test scores." *Economic Inquiry* 46 (1), pp. 2-12.

Borghans, L., B. Golsteyn, J. J. Heckman, and J. E. Humphries (2011), "IQ, Achievement and Personality." *Personality and Individual Differences* 51, pp. 315–320.

Borghans, L. and T. Schils (2012), The leaning tower of Pisa: decomposing achievement test scores into cognitive and noncognitive components, Paper presented at the 2012 NBER Summer Institute in Education Economics, Cambridge.

Borghans, L. and T. Schils (2012), How much should we test at schools? About labor market signals and learning incentives, Paper presented at the 2012 SOLE Conference, Chicago.

Duckworth, A.L., P.D. Quinn, D. Lynam, R. Loeber, T. Moffit, and A. Caspi (2010), "What intelligence tests test: Individual differences in test motivation and IQ." Unpublished manuscript, University of Pennsylvania.

Lazear, E. (2006), Speeding terrorism and teaching to the test, *The Quarterly Journal of Economics*, 121 (3): 1029-1061.

Neal, D. and D. Whitmore Schanzenbach (2007), Left behind by design: proficiency counts and test-based accountability, NBER WP no. 13293, Cambridge.

Segal, C. (2011), "Working When No One is Watching: Motivation, Test Scores, and Economic Success" forthcoming in *Management Science.*