pedocs
Open Access Erziehungswissenschaften

DIPF

Penk, Christiane; Richter, Dirk

Change in test-taking motivation and its relationship to test performance in low-stakes assessments

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Educational assessment, evaluation and accountability 29 (2017) 1, S. 55-79

Bitte verwenden Sie in der Quellenangabe folgende URN oder DOI /
Please use the following URN or DOI for reference:
urn:nbn:de:0111-pedocs-174284
10.25656/01:17428

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-174284
https://doi.org/10.25656/01:17428

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document.
This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Mitglied der Leibniz-Gemeinschaft

Change in Test-Taking Motivation and Its Relationship to Test Performance in Large-Scale Assessments

Abstract

Since the turn of the century, an increasing number of low-stakes assessments (i.e., assessments without direct consequences for the test-takers) are being used to evaluate the quality of educational systems. Internationally, research has shown that low-stakes test results can be biased due to students' low test-taking motivation, and that students' effort levels can vary throughout a testing session involving both cognitive and noncognitive tests. Thus, it is possible that students' motivation vary throughout a single cognitive test and in turn affect test performance. This study examines the change in test-taking motivation within a two-hour cognitive low-stakes test and its association with test performance. Based on expectancy-value theory, we assessed three components of test-taking motivation (expectancy for success, value, and effort) and investigated its change. Using data from a large-scale student achievement study of German ninth-graders, we employed second-order latent growth modeling and structural equation modeling to predict test performance in mathematics. On average, students' effort and perceived value of the test decreased, whereas expectancy for success remained stable. Overall, initial test-taking motivation was a better predictor of test performance than change in motivation. Only the variability of change in the expectancy component was positively related to test performance. The theoretical and practical implications for test practitioners are discussed.

Keywords: test-taking motivation, low-stakes tests, large-scale assessments, expectancy-value theory, growth modeling

Acknowledgments: We thank Sara J. Finney for her enriching comments and methodological support as well as Bo Bashkov for proofreading the manuscript

Introduction

Students approach tests with different attitudes and affects and thus may engage in different behaviors. Tests with short- or long-term consequences for the students are called *high-stakes* tests. Given the consequences of such tests, practitioners seem to assume that students exert much effort during high-stakes tests due to the personal consequences for them. However, since the beginning of the Programme for International Student Assessment (PISA) in the year 2000, assessments without consequences for the test-takers (i.e., *low-stakes* tests) have become increasingly important for evaluating the quality of Germany's educational system (Stanat & Lüdtke, 2013). Test-taking motivation (TTM) is an important issue under these circumstances, because it is possible that students do not give their best effort due to the lack of any personal consequences of the test results.

TTM refers to students' readiness to engage in completing the test and became a growing research area of interest in the past decade. Several studies have shown that motivated test-takers outperform unmotivated test-takers (Baumert & Demmrich, 2001; Cole, Bergin, & Whittaker, 2008; Eklöf, Pavešič, & Grønmo, 2013; Thelk, Sundre, Horst, & Finney, 2009, Wise & DeMars, 2005). Thus, the impact of motivation on performance is very important for the interpretation of test results in low-stakes assessments. If the students do not give their best effort, it remains unclear whether test scores correspond to the true ability of the students.

In addition to the issue of initial motivation with which students approach the testing session, it is also important to consider the potential *change* in motivation throughout a long testing session and its effect on students' test performance. TTM can both increase and decrease throughout a test. An increase in TTM can be interpreted as flow (Moneta & Csikszentmihalyi, 1996). A decrease in TTM can be interpreted as fatigue or boredom (Barry & Finney, 2016). However, in a low-stakes testing context it is more conceivable that the

students show a decrease in motivation within a cognitive (mentally taxing) test because TTM depends for example on the item difficulty (Wise & Smith, 2011), item location (Wise, 2006), or the item type (Sundre & Kitsantas, 2004). A study investigating the change in motivation found no fatigue effect (i.e., a decline in effort) during a low-stakes testing session with different test types, i.e., cognitive and noncognitive test (e.g., measures of attitudes). However, it did find that TTM was influenced by test-specific characteristics, such as mental taxation (Barry & Finney, 2016). Studies exploring the change in TTM within one cognitive test found a decrease in TTM (Horst, 2010; Wise, 2006; Wise, Pastor, & Kong, 2009), but they did not investigate the relationship between the change in TTM during the cognitive test and test performance. This relationship, however, is of particular interest for large-scale assessments evaluating the outcomes of educational systems. Using the results of low-stakes tests without knowing the effect of change in TTM on test results can threaten the validity of inferences based on those test results (Eklöf, 2008, 2010a; Thelk et al., 2009). Thus, the current study aims to investigate (a) the change in TTM based on the expectancy-value theory (EVT) and (b) the relationship between change in TTM and test performance in a German low-stakes large-scale assessment. This investigation is especially important for several reasons: a) change in effort during a test is often discussed but rarely empirically evaluated in research dealing with TTM; moreover, when it is studied, change in effort is not linked to actual test performance; b) the expectancy component is often ignored in applications of the EVT; and c) the role of effort as a mediator variable between performance and both the expectancy and value components requires further investigation. Each of these issues is explained in more detail below. In the next section, we first define the construct of TTM and integrate it into the framework of EVT. We then provide an overview of previous research.

Expectancy-Value Theory and the Non-Longitudinal Assessment of TTM

EVT is a frequently used framework in the context of TTM (Eccles & Wigfield, 2002; Sundre, 2007; Wigfield & Eccles, 2000). As shown in Figure 1, EVT assumes that the expectancies for success and the perceived value of a test directly affect achievement behavior, which involves both the expended effort on the test and actual test performance. Expectancies refer to students' perceptions of how well they will perform. The value component includes four distinct aspects: attainment value (the importance of the test), intrinsic value (the enjoyment during the test), utility value (the usefulness of the test), and cost (e.g., test anxiety).

Insert Figure 1 about here

Test-taking motivation is defined as "the willingness to engage in working on test items and to invest effort and persistence in this undertaking" (Baumert & Demmrich, 2001, p. 441). Thus, test-taking effort constitutes the main element of TTM and is described as the engagement of the test-takers and their expenditure of energy to achieve the best possible test score (Wise & DeMars, 2005). According to EVT, effort is the outcome of expectancy and value, and is therefore related to test performance. This means that effort should mediate the relationship between performance on the one hand and expectancy and value on the other hand. In sum, TTM construct includes all three components: the effort that the students invest, their expectancy for success, and the value they place in the test.

Typically in school settings, the expectancy component has been shown to be a stronger predictor of test performance than the value component, which was more closely associated with persistence or task choice (Pekrun, Elliot, & Maier, 2009; Schunk, Pintrich, & Meece, 2008; Wigfield, 1994; Wigfield & Eccles, 2000). However, in low-stakes tests, most

research only considers the value component and effort when examining TTM (Eklöf & Nyroos, 2013; Eklöf et al., 2013; Wolf & Smith, 1995) and ignores the expectancy component, "because test-takers in low stakes tests seldom have any way of finding out if they were successful" (Cole et al., 2008, p. 613). Overall, studies have found a positive relationship between the value component and test performance, as well as between effort and test performance. Cole et al. (2008) as well as Zilberberg, Finney, Marsh, and Anderson (2014) have investigated and found support for the mediating role of effort. The few studies that included both the expectancy and value components found that expectancy for success predicted test performance in low-stakes tests (Asseburg, 2011; Freund & Holling, 2011); however, these studies did not examine effort. In sum, most of the studies ignored at least one aspect of TTM or the mediating role of test-taking effort. Although the EVT was not developed to model the change in motivational processes we adapt the model accordingly and investigate the relationship between change in expectancy, change in value, and change in effort.

Longitudinal Assessment of TTM

The longitudinal assessment (i.e., the change in TTM) of motivation is important, because achievement tests often take much longer than a regular school period (45 minutes). Therefore, a long testing session may lead to fatigue and a decrease in TTM (Cao & Stokes, 2008). According to EVT, it is also conceivable that a loss of high expectancy for success during the test can result in a decline in effort, resulting in low test performance. In particular, due to recurring difficult items that the test-takers cannot solve might change their confidence to answer the next items and their willingness to invest effort. This dynamic was uncovered by Wise and Smith (2011) in their demands-capacity model of test-taking effort that includes aspects of initial effort and potential change in effort during the test. The authors emphasized

the dynamics of TTM to make a meaningful interpretation of the test results. However, very few studies consider the dynamic of TTM in their analyses.

Some studies have examined these dynamics or the change in TTM over a low-stakes testing session using response behavior at the item level in computer based assessments (Wise, 2006; Wise et al., 2009). The studies found that students guessed the correct response more often if the item was located in the back of the test booklet. This indicates that TTM may decrease throughout the test, for instance, due to fatigue or change in confidence in the ability to answer future items correctly (Wise & Smith, 2011). At present, however, most large-scale assessments are traditional paper-and-pencil tests and one cannot use an electronically recorded measure of TTM. For this type of assessment one can only employ self-reported measures of motivation that could also serve as valid indicators of test-taking motivation in low-stakes tests (Swerdzewski, Harmes, & Finney, 2011).

Thus, other studies have examined change in TTM on paper-and-pencil tests (Barry & Finney, 2016; Barry et al., 2010; Horst, 2010). These two studies focused on three hour long testing sessions including one cognitive and four noncognitive tests and assessed students' effort and perceived importance of the test after completing each of the five tests. In sum, students reported less effort on the cognitive test and more effort on the noncognitive tests independent of the order in which the tests were administered. That suggested that in low-stakes assessments students are less willing to invest effort in mentally taxing tests. Moreover, Barry and Finney (2016) investigated the *change* in effort and importance across one cognitive and four noncognitive tests. In general, effort slightly increased during the testing session, with the smallest reported effort score found for the cognitive test that was administered first. Barry and Finney assumed that the rise in effort over the testing session is probably due to the low mental taxation of the noncognitive test in relation to the high mental taxation of the cognitive test (i.e., higher cost in light of EVT). Thus, students' effort may

decrease within one high mental taxing test. In contrast, students rated the cognitive test as the most important, even though they invested the least effort in it in comparison to the other tests. Effort and importance were moderately correlated for the cognitive test, but the *change* in effort and the *change* in importance in within all five tests were not. However, it is possible that the change in effort and change in importance within a *cognitive test* is related. Barry and Finney (2016) also assessed the expectancy component (i.e., self-efficacy) after the students completed all tests, but it was not related to students' effort on either the cognitive or noncognitive tests. Importantly, this measure was collected only once for the cognitive test, so no conclusion about the change in expectancy can be drawn.

Horst (2010) used a design similar to that of Barry and Finney. She split the cognitive test administered at the beginning of the assessment into three subtests to assess how TTM changes *within a cognitive test*. Although the test was viewed as important the entire time, the reported effort decreased, probably due to the high "cost" of the cognitive test. It is possible that the student demonstrated a lower level of test performance due to the diminished level of effort than they could have demonstrated with a stable level of effort. Additionally, effort and importance showed higher relationships for the cognitive test than for the noncognitive tests, indicating that the two components of EVT are more closely associated for cognitive tests.

To summarize, only a few studies (Horst, 2010; Wise, 2006; Wise et al., 2009) have investigated the change in TTM within a cognitive test; they found evidence for a small fatigue effect. Additionally, only one study (Barry & Finney, 2016) has assessed the expectancy component; however, this study did not assess the change in expectancy for success throughout the testing session. Furthermore, no study has investigated the relationship between change in TTM within a cognitive test and test performance. This is especially important for large-scale assessments, because changes in TTM may impact the results of the

tests and therefore limit or even bias their interpretation. Test scores from test-takers who show a decrease in TTM are likely to underestimate actual ability or achievement.

Study Objectives and Research Questions

The purpose of this study is to build upon previous research (Barry & Finney, 2016; Horst, 2010) by modeling the *change in all three components of EVT* (expectancy, value, and effort) and relating the change in these constructs to test performance. TTM was assessed within a two-hour low-stakes assessment at three measurement points: before the test, after half of the test, and after the test. Thus, this study is similar to the one conducted by Horst (2010) in that TTM was also assessed throughout *a* cognitive test. Unlike to Horst (2010), however, this study also assessed initial TTM measured before the test and change in the expectancy component. Altogether, we focus on two main quantities of change: a) average change and b) variability in intra-individual change. The average change (a) describes the mean rate of change of *all* students within the testing session. The variability in intra-individual change (b) refers to the *individual* variation in change because the individual test-takers have different trajectories (i.e., individual differences in intra-individual change): For example, some test-takers show a decline in effort, some an increase and some either a decline or an increase in effort. The next subsections describe the research questions in more detail and point out to the type of change we focus on.

Change in Test-Taking Motivation

(1) Does TTM change on average within a two-hour cognitive low-stakes testing session?

The first question pertains to the average change (a) in expectancy for success, perceived value of the test, and test-taking effort. Based on previous research that found a slight decrease in effort within one cognitive test and a stable level of perceived importance of the test (Cao & Stokes, 2008; Horst, 2010), we expect at least a slight decrease in effort and

essentially no change in perceived importance (i.e., the attainment value of EVT should be stable). Although Barry and Finney (2016) found an increase in effort during the testing session, we still expect to see a decrease in effort because similar to Horst (2010) we explore change in effort within a single cognitive test.

Previous research did not even consider the change in expectancy for success within a testing session. We assume that there are two possibilities regarding the change in expectancy. The first possibility supposes that, on average, students' expectancy for success should not change much across the three time points. Based on EVT, the average level of expectancy for success should remain stable because students who know the domain in which they are being tested can also estimate their corresponding competence level. Thus, students should not change much in expectancy throughout the testing session (little intraindividual change). As such, the average expectancy across test-takers at the three time points should be approximately the same. The second possibility assumes the presence of change in expectancy for success within a test. According to the demands-capacity model of test-taking effort (Wise & Smith, 2011), the expectancy for success could change due to the completion of previous test items. Given individual differences in students' expectancies, there could also be interindividual differences in change across test-takers (some increasing in expectancy, some decreasing in expectancy, some staying the same, as described in b). However, when averaged over test-takers (a), such differences manifest as three averages that are very similar in magnitude (because some test-takers increase, whereas others decrease and yet others stay the same). Due to the two theoretical possibilities described above, we cannot form a hypothesis about the change in expectancy.

(2) Is the change in some TTM constructs related to the change in other TTM constructs?

Previous research has found no relationship between the change in effort and the change in value or between expectancy and change in effort in a testing session with different test types (Barry & Finney, 2016). Moreover, previous research has shown that effort and value were related to each other on a cognitive test (Barry & Finney, 2016). Because this study focused on the change of TTM within one cognitive test with a constant level of mental taxation (in contrast to a testing session including also noncognitive tests) we assume that the change in effort and the change in value are related. Although we expect a stable level of value on average (a; see research question 1), it is possible that students show different trajectories, similar to the second hypothesis for the change in expectancy (b). Based on the demands-capacity model of test-taking effort, that relates the change in confidence to solve future items (i.e., the change in expectancy) with the change in effort we assume that the change in expectancy for success and the change in effort are related. To understand the mechanism driving these relationships between the TTM constructs it is important to examine not only change on average, but individual change trajectories.

Change in TTM and Its Relationship to Test Performance

(3) What is the relationship between change in TTM and test performance after accounting for students' socio-demographic background, ability in mathematics, and domain-specific motivation?

According to EVT, expectancy and value influence the expended effort on a given test. It is reasonable to assume that students with a decreasing level of TTM show a lower test performance relative to students with a stable level of TTM. In particular, as the level of effort declines, students may answer only easy items or abandon the test altogether, which would manifest in a low test score. In contrast, change in effort may not be related to test performance because those students who decrease in effort over time may still score higher or

lower on the total test than those students who remain stable in effort over time. Thus, the relationship between change in effort and total test performance is difficult to interpret.

Method

Sample

The current study explores TTM in the German *National Assessment Study* conducted by the Institute xxx [blinded for review]. This study is a typical low-stakes assessment, and it measured mathematical and scientific literacy in a representative sample of German ninth-graders (N = 44,584). Students with special needs (N = 1,380) were excluded because they received a different test from the rest of the sample. In addition, students were excluded if they intentionally disregarded the instructions of the TTM questionnaire (i.e., handing in a completely blank questionnaire or providing the same response option on all items; N = 906). Thus, a total of 42,298 students were included in this study. About half the sample (49.8%) was female, and the mean age was 15.6 years (SD = 0.61). One third of the students (35.2%) attended the academic-track; the remaining students were assessed at non-academic track schools.

Procedure and Instruments

Procedure. The assessment took place in spring 2012, and TTM was assessed at three measurement points. The TTM items were presented for the first time after the general instructions for the test (T1). Following the TTM questionnaire, students worked on the first half of the achievement test. After the first test half (one hour) and a 15 minute break, students completed the TTM items again (T2). After finishing the second test half and taking another break, students completed a socio-demographic background questionnaire. For approximately half of the sample, this questionnaire contained items on TTM (T3). In summary, during the two-hour testing session, TTM was assessed three times: before the test, after the first half of the test, and (for half of the sample) after the test.

Achievement test. The test is a standards-based assessment designed to evaluate and compare mathematical and scientific competencies of students in the German federal states (*National Assessment Study 2012;* Pant et al., 2013). Our study focused only on mathematics. A balanced incomplete block design was used (i.e., every student was administered only a subset of all items; Frey, Hartig, & Rupp, 2009), in order to administer a sufficient number of items of the test domain within a limited testing period. Weighted likelihood estimates (Warm, 1989) were computed as measures of test performance in mathematics. These estimates are based on unidimensional scaling with the 1-parameter logistic (Rasch) model. The test was a typical low-stakes assessment for the test-takers because they did not receive a grade or individual feedback on their test performance. The mathematics test showed a high reliability (WLE Person separation reliability = .91; EAP/PV reliability = .91).

Test-taking motivation. We used the Questionnaire on Current Motivation to measure both the expectancy and value components of EVT (Freund, Kuhn, & Holling, 2011; Rheinberg, Vollmeyer, & Burns, 2001). The scale of the items ranged from 1 = strongly disagree to 4 = strongly agree. The expectancy component is represented by the perceived probability of success (e.g., "I think I am up to the difficulty of this test"), and the value component is represented by the challenge subscale (e.g., "I am eager to see how I will perform on the test"). The concept of challenge denotes the extent to which test-takers perceive the situation as an achievement situation and corresponds to the attainment value in EVT (i.e., perceived importance of the test). The challenge subscale is henceforth referred to as the importance subscale. At the beginning, the questionnaire assesses the current motivation before taking the cognitive test. Then, the items were adapted to assess TTM after the first half of the test and again after the test. Additionally, test-taking effort was assessed with items from the TTM scale by Eklöf (2010b; e.g., "I worked on each item in the test and persisted even when the task seemed difficult") with a scale ranging from 1 = strongly

disagree to 4 = strongly agree. All scales relate to the current test situation and aim to measure states (as opposed to traits). The effort scale was originally developed to measure invested effort after a test, but these items were also adapted for the first and second time point.

The measurement model for all five TTM constructs was established by Penk and Schipolowski (2015). Thus, our study used a four-item effort scale, a three-item expectancy scale, and a two-item importance scale (for more information see Penk & Schipolowski. 2015). The descriptive statistics of the TTM subscales are displayed in Table 1. The values of the three subscales indicated a decrease in effort and importance during the test. In contrast, probability of success appeared to be stable. The reliability estimates for the factors were acceptable given the small number of items on each scale.

Insert Table 1 about here

Student Background Questionnaire. Students completed an instrument with self-report scales at the end of the test, e.g. questions on the student's home environment and self-related beliefs. In this study, we used the data for students' self-concepts in mathematics as a measure of domain-specific motivation. This construct was measured with four items (e.g., "I get good grades in mathematics"; Ramm et al., 2006) on a four-point Likert-scale ranging from 1 = strongly disagree to 4 = strongly agree (McDonald's ω= .91). Five variables from the student background questionnaire were used as control variables: (a) sex, (b) school track, (c) socio-economic status, (d) immigration background, and (e) grade in mathematics. Previous research has found differences in TTM between academic-track students and nonacademic-track students, so we included the school track as a control variable in our analyses (Penk, Pöhlmann, & Roppelt, 2014). Socio-economic status was assessed with the Highest International Socio-Economic Index of Occupational Status (HISEI; Ganzeboom, De Graaf,

& Treiman, 1992), which is an indicator of the status of the parents' professions with respect to income and education level. The HISEI scores were standardized. Students' immigration background was defined in the following way: (a) one parent was not born in Germany, (b) both parents were not born in Germany, but the student was born in Germany, or (c) both parents and the student were not born in Germany (Stanat & Christensen, 2006). Students' grade in mathematics (standardized and centered at their class mean) was included to control for students' relative ability in mathematics before taking the test. Due to Germany's grading system lower values indicate a higher ability level than higher values.

Analyses

In this study, second-order latent growth curve modeling was used to examine the trajectories of the TTM constructs within a test. This modeling framework allows for estimating the initial level of TTM as well as the change in TTM at three measurement points. It makes it possible to examine intraindividual change in expectancy, value, and effort during the test, as well as interindividual differences in this intraindividual change (Sayer & Cumsille, 2001). S*econd-order* latent growth curve models use latent variables to estimate growth over time (e.g., a latent motivation variable composed of four effort items measured at three time points). Thus, the latent variables form the *first-order factors*, and the growth parameters form the *second-order factors* (Ferrer, Balluerka, & Widaman, 2008; Sayer & Cumsille, 2001). See Figure 2. This technique allows for the separation of measurement error from true trait change and reliable time-specific variance. In addition, these models also make it possible to test the assumption of measurement invariance over time. Finally, they have the statistical power needed to uncover individual differences in change (Geiser, Keller, & Lockhart, 2013; Sayer & Cumsille, 2001).

Knowing the advantages of second-order latent growth curve modeling we describe now the model specification. Latent growth curve modeling includes an intercept and one or

more slope factors as growth factors. Due to the three measurement points in this study, we could only estimate linear trajectories (i.e., one linear slope factor) and could not test more complex shapes of growth, such as quadratic or piecewise growth (Hancock, Kuo, & Lawrence, 2001). The intercept factor represents the initial level of the variable of interest (e.g., effort before the test). As shown in Figure 2, the paths from the intercepts (e.g., initial effort) to the three latent first-order factors (e.g., effort variables at three time points) are fixed at 1, because the intercept is a stable constant without growth. The slope factors describe the linear rate at which the variable of interest changes over time, e.g., a decrease in effort during the testing session (Preacher, Wichmann, MacCallum, & Briggs, 2008). Hence, the paths from the slope factor (e.g., change in effort) to the three latent first-order factors (e.g., effort variables at three time points) are fixed at 0, 1, and 2, respectively, reflecting the linear trajectories. The path from the slope factor to the first time point is fixed to zero because there can be no growth at the initial time point.

Change in test-taking motivation. Three linear growth models are estimated to answer the first research question investigating change in TTM during the testing session: one model each for probability of success, perceived importance, and test-taking effort. All covariances among the first-order factors (e.g., covariances among the latent effort variables at the three time points) were set to zero under the assumption that the relations among the first-order factors are explained fully by the second-order latent growth factors (Sayer & Cumsille, 2001). For the second research question examining whether change in the three TTM constructs is related to each other, we simultaneously estimated the growth processes of probability of success, importance, and effort using one multivariate second-order latent growth model. This model allows correlations among all of the growth parameters of the three TTM constructs. Thus, all of the intercept and slope factors for probability of success, importance, and effort were estimated in one model and were allowed to correlate.

Change in TTM and its relationship to test performance. For the last research question, we used a two-step procedure estimating two consecutive models. We first examined how much variance in the test scores could be explained by students' socio-demographic background (i.e., sex, school track, socio-economic status, and immigration background) variables (Model 1). In the next step, we predicted test performance with the growth parameters of TTM within the testing session. Thus, we added the intercept and slope factors for expectancy, importance, and effort as predictors of test performance in mathematics. The self-concept in mathematics was added as a predictor of both test performance and probability of success to control for any spurious relationship between the two constructs due to both being related to self-concept (Eccles & Wigfield, 2002; Eklöf, 2006, 2007, 2008). The intercepts and slopes for perceived importance and probability of success were used as predictors of the corresponding growth factor of effort. In this way, we want to test primarily whether the effect of change in importance and change in probability of success on test performance is mediated by change in effort (Cole et al., 2008; Zilberberg et al., 2014). The final model explores how much of the variance in test performance is associated with the state-like TTM constructs, while controlling for the students' socio-economic background characteristics and domain-specific motivation.

Insert Figure 2 about here

Analyses were conducted in M*plus* 7.1; (Muthén & Muthén, 2012). The hierarchical structure of the data (i.e., students nested within classes) was taken into account in the computation of standard errors and model fit. In addition, sample weights were used in all of the analyses to ensure the results are representative of the population of ninth-graders in German schools. Due to the large sample size (i.e., high power), we specified a *p*-value

below .001 as the cut-off for statistical significance. In all of the analyses, we applied a robust maximum likelihood estimator and considered the following indices to evaluate the model fit: (a) MLR χ^2-statistic, corresponding degrees of freedom, and probability value; (b) comparative fit index (CFI); (c) Tucker-Lewis index (TLI); (d) root mean square error of approximation (RMSEA); and (e) standardized root mean square residual (SRMR). According to Hu and Bentler (1999), the following values indicate adequate model fit: CFI > .95, TLI < .95, RMSEA < .06, and SRMR < .08. Strong measurement invariance is required to apply second-order latent growth curve modeling. This requirement ensures that the items in the questionnaire assess the same construct at every measurement point. To compare the different measurement invariance models (configural, metric, and strong measurement invariance), we used ΔCFI (Cheung & Rensvold, 2002; Rutkowski & Svetina, 2014) and the Root Deterioration per Restriction statistic (RDR; Browne & Du Toit, 1992), comparing the relative fit of nested models based on their RMSEA differences. Values of ΔCFI < -.01 and values of RDR < .05 suggested a good model fit. As mentioned by Barry and Finney (2016), although we apply these general guidelines to evaluate the structural equation models, one should keep in mind that there is currently little research on applying these indices to second-order latent growth models.

Results

Before presenting the results, it is important to test whether the data meet the requirements for the application of second-order latent growth modeling. The model requires the assumption of strong measurement invariance. More specifically, the constructs of interest need to be represented by the same structure over time (i.e., the construct consists of the same indicators over time; configural measurement invariance), the same numerical factor loadings of each indicator over time (metric measurement invariance), and the same numerical intercepts for each indicator over time (strong measurement invariance). Strict measurement invariance

(indicators have the same error variances over time) is generally unlikely in growth models due to heterogeneous variance over time (Sayer & Cumsille, 2001). Appendix A includes the results of these nested measurement invariance models. The constructs used in this study exhibited strong measurement invariance (effort: $\chi^2(51) = 743.16$, $p < .001$; CFI = .99; TLI = .99; RMSEA = .02; SRMR = .03; RDR = .03; probability of success: $\chi^2(23) = 265.89$, $p <$ .1 ; CFI = .99; TLI = .99; RMSEA = .02; SRMR = .03; RDR = .03). The importance scale consists of only two items. As invariance tests require at least three indicators, we cannot use goodness-of-fit statistics to evaluate model fit (Bollen, 1989). However, the second-order latent growth model assuming strong measurement invariance for the importance construct fits the data well ($\chi^2(7) = 547.12$, $p < .001$; CFI = .98; TLI = .97; RMSEA = .04; SRMR = .03). This suggests that the importance factor also exhibits strong measurement invariance.

Change in Test-Taking Motivation

The first research question addressed change in TTM over a two-hour low-stakes testing session. As shown in Table 2, all three second-order latent growth curve models estimating the linear change in effort, importance, and probability of success showed satisfactory fit. The mean of the effort intercept was 2.98 on a 4-point scale, indicating that, on average, students reported that they were willing to invest effort before the test. The coefficients in Table 2 differ slightly compared to the coefficients in Table 1 because the former are latent values and the latter manifest values. The mean of the linear slope was -0.13 (*beta* = -.81), indicating that students' average effort decreased during the testing session as hypothesized. After half of the test, the mean average effort decreased to 2.84 [≈ 2.98 + (1 × -0.13)], and at the end of the test, the mean average effort was 2.71 [≈ 2.98 + (2 × -0.13)]. That means effort decreased with more than 0.5 standard deviation of the effort intercept (*SD* = 0.47), which is a moderate decline statistically as well as practically. The variances of the intercept and slope factors

were statistically significant, but students showed much more variability in their initial effort than in their change in effort. Assuming that the growth parameters followed a normal distribution, the estimated means and variances can be used to generate a distribution of the change in effort. Approximately two-thirds of the students showed slope values between -0.30 and 0.03; that is, some showed a greater decrease in effort, whereas others remained more or less stable. The correlation between the intercept and slope for effort was negative and statistically significant (r = -.14), indicating that students with a higher initial effort than average had a greater decrease in effort than average. However, the correlation was quite small.

The growth parameter estimates for the importance factor were similar to the estimates for effort. The mean of the importance intercept was 2.70, indicating that, on average, students perceived the test as important before they took the test. The mean of the importance slope was negative and statistically significantly -0.18 (*beta* = -.73). In other words, contrary to our hypothesis the level of importance decreased significantly during the testing session. Again, the variances of the intercept and slope factors were significant, and students showed greater variability in their initial importance than in change in importance. The variability in importance, especially before the test, was very high, indicating students valued the test quite differently from one another. However, about two-thirds of the students demonstrated slope factors ranging from -0.43 to 0.07, which indicated that some students perceived the test as important during the entire test, whereas for others importance decreased more than the average decrease. In contrast to the effort growth parameters, the importance intercept and slope were not significantly correlated.

Insert Table 2 about here

The initial perceived probability of success was 2.89, which is similar to the initial levels of the other two constructs. Prior to the test, the average student felt confident that they would complete the test successfully. However, the mean of the slope of probability of success was almost 0 (*beta* = -.11), though statistically significant, indicating that students' probability of success remained mostly stable. The variances of the intercept and slope factors were significantly different from zero, and again students showed more variability in the initial probability of success than in change in the probability of success. Considering the standard deviation, approximately two-thirds of the students had a mean slope ranging between -0.21 and 0.16, indicating that some students reported a decrease in their perceived probability of success, whereas others reported an increase. The intercept and slope for probability of success were not correlated.

The second research question focused on the relationship between the growth parameters for effort, importance, and probability of success. For this purpose, we modeled the three growth processes simultaneously in one multivariate second-order latent growth curve model. Table 3 contains the factor correlations. The estimated model fitted the data well ($\chi^2(311) = 7781.83$, $p < .001$; CFI = .96; TLI = .95; RMSEA = .02; SRMR = .06). All of the slope factors were positively correlated with each other. The intercept of the effort factor and the intercept of the importance factor showed the highest correlation, as did the slope of the effort factor and the slope of the importance factor ($r = .79$ for both). As both slopes were negative, the correlation expresses that a smaller decrease in importance for students is associated with a smaller decrease in test-taking effort over the testing session. In other words, students who decrease more in effort relative to other students tend to decrease more in importance relative to other students. The correlation between the slopes for probability of success and effort was also significant ($r = .45$) and indicated that a smaller decrease in probability for success tends to be accompanied by a smaller decrease in test-taking effort. Additionally, the slopes for importance and probability of success were moderately correlated

(r = .33), indicating that the smaller the decrease in probability of success, the smaller the decrease in importance. Moreover, the intercepts for effort and probability of success showed a small correlation (r = .21), indicating that test-takers who decreased more than average on probability of success tended to decrease more than average on effort.

Insert Table 3 about here

To sum up the first two research questions, the three second-order latent growth curve models showed a moderate initial TTM before the test and a moderate decrease in effort and importance within the test. In contrast, students' probability of success remained stable. The slopes of the three TTM constructs were significantly correlated with each other, indicating moderate to strong relationships between the changes in effort, importance, and probability of success.

Change in TTM and Its Relationship to Test Performance

The last research question investigated the change in TTM and its relationship to test performance in a two-step procedure. First, we predicted the mathematics score solely with the student's background information: sex, school track, socio-economic status, immigration background, and grade in mathematics (Model 1). This model did not include any motivational variables. In the second step, all of the growth parameters of the three TTM constructs as well as the domain-specific motivation were added as predictors of test performance (Model 2).

The first step of the procedure is analyzed in Model 1. The model showed good fit ($\chi^2(25) = 169.96$, $p < .001$; CFI = .99; TLI = .99; RMSEA = .02; SRMR = .01). The five background variables significantly predicted the mathematics scores and explained 57% of

their variance. The strongest predictor was school track (*beta* = .59), indicating that students attending the academic track outperformed their classmates in non-academic tracks. The grade in mathematics (*beta* = -.33) also predicted students' test performance. Sex (*beta* = .12), immigration background (*beta* = -.12), and socio-economic status (*beta* = .10) also significantly predicted test performance, but these coefficients were quite small. Specifically, male students outperformed female students, students without an immigration background outperformed students with an immigration background, and the higher the socio-economic status, the higher the test score of the student.

The second model (presented in Figure 3) also fit the data well. The three TTM constructs, students' background variables, their grade in mathematics, and their domain-specific motivation explained 64% of the variance in mathematics scores. The latent growth curve models of the three TTM constructs and the domain-specific motivation explained an additional 7% of the test score variance (1% trace back to self-concept in mathematics).

Insert Figure 3 about here

Looking at the paths in the model in more detail, self-concept in mathematics was a predictor of test performance as well as a strong predictor of the probability of success intercept. Specifically, self-concept in mathematics explained almost a quarter of the variance in initial probability of success, but did not predict the change in probability of success. After controlling for self-concept in mathematics, the probability of success intercept and slope significantly predicted test performance; the higher the initial probability of success and the smaller the decrease in probability of success (or the greater the increase), the better the student's test performance.

The intercept and slope for importance had no significant direct effects on test performance, but they significantly predicted the respective effort factors. Over 60% of the variance in the growth factors for effort could be explained by the growth factors for importance and probability of success. However, the intercept and slope for effort were mainly predicted by the intercept and slope for importance. In addition, the intercept for effort significantly predicted test performance. Thus, the higher the initial effort, the better the student's performance.

Of the indirect effects shown in Appendix B, only the effect of the intercept for importance on test performance via effort was substantial and significant (*beta* = .13). Thus, the effect of initial importance on test performance was fully mediated by the initial level of effort. The indirect effect of the initial probability of success on test performance was significant but quite small (*beta* = .03). Moreover, the intercept and slope for probability of success showed significant and substantial direct effects on test performance. Neither was mediated by effort.

In summary, the probability of success intercept and slope and the effort intercept were direct predictors of the mathematics scores after accounting for students' background characteristics. The importance intercept and slope were not directly related to test performance, but they directly predicted the test-taking effort intercept and slope, respectively. Effort mediated only the effect of the importance intercept on test performance. Both the intercept and slope for probability of success predicted test performance after controlling for domain-specific motivation. Thus, there appears to be a relationship between the state variable probability of success and test performance, beyond the trait-like variable self-concept in mathematics. Additionally, the test-taking effort intercept was the strongest predictor of test performance among the parameters of the TTM constructs. The sizes and, therefore, the effects of the significant growth parameter coefficients on test performance

were comparable to the background variables: socio-economic status and immigration background.

Discussion

As the number of low-stakes tests in German schools has increased, research on test-taking motivation (TTM) has grown in the last decade. Research shows that test scores from low-stakes assessments may be affected by low motivation (Cole et al., 2008; Eklöf & Nyroos, 2013; Swerdzewski et al., 2011; Wise & DeMars, 2005; Wolf & Smith, 1995). Understanding the mechanism of TTM during the testing session and the effects of TTM on the test scores is crucial to ensure the proper interpretation of test results (Thelk et al., 2009). Thus, the current study had two main purposes. First, we explored the change in three TTM constructs based on expectancy-value theory (EVT). Specifically, we examined probability of success (expectancy), perceived importance of the test (attainment value), and test-taking effort within a testing session in which students completed a cognitive test in mathematics. Our second goal was to investigate the relationship between change in the three TTM constructs and students' test performance.

Change in Test-Taking Motivation

The first research question addressed the average change in probability of success, perceived importance of the test, and test-taking effort. The results showed that, on average, probability of success remained stable over the testing session, but perceived importance of the test and test-taking effort decreased within the testing session. Despite the significant variability in the change in all three TTM constructs, it can be considered "good news" that, on average, students reported a moderate decrease on two of the three TTM constructs, although they completed a two-hour cognitive (and mentally demanding) test without any personal consequences. The moderate decrease might be due to the break students had to recover from the first half of the test. It seems that a two-hour low-stakes test is an adequate time frame,

considering intraindividual motivational processes. The results of our study are in line with Horst's (2010) findings, which demonstrated a slight decrease in test-taking effort on one cognitive test (d = 0.19) over a 50-minute period. The students in our study reported a slightly larger change in the effort factor ($d_{T1/T2}$ = 0.31; $d_{T2/T3}$ = 0.23) than the students in Horst's study that did not include a measure of TTM before the test, as we did in our study. However, on average, probability of success remained at a stable level. Thus, although students were asked before the test, it appears that they provided realistic estimates of their expectancy for success. In contrast to Horst's results, the importance scale in our study showed a decrease similar to that of effort. This may be attributable to the fact that we measured the attainment value of EVT indirectly using the challenge scale, instead of directly measuring the perceived importance of the test.

Overall, test-takers showed more variability in their initial TTM than in their change in TTM. The variability in initial importance was especially high, indicating that students varied a lot in their perceived value of this test. Although beginning at different levels of TTM, on average students showed similar change in TTM throughout the test. Thus far, no other study has investigated change in the value and expectancy components within one cognitive test, so we cannot compare the findings with those of other investigations. However, Horst's (2010) results indicated a fairly stable level for importance and lower variances for the three importance means in comparison to the effort means. The variability of change in the TTM constructs was small, but nevertheless significantly different from zero.

Before investigating the relationship between change in TTM and test performance, we explored whether the changes in probability of success, perceived importance of the test, and test-taking effort were related to each other. The results showed a strong relationship between initial importance and initial effort, as well as between the change in importance and change in effort. If students valued the test and retained this attitude throughout the test, they

were more willing to invest effort throughout the entire testing session. Moreover, initial probability of success was related to initial effort. Although we found that, on average, level of probability of success was stable throughout the test, change in probability of success was related to change in effort. Thus, it appears that students' individual trajectories vary enough for change in probability of success to correlate with change in effort. Effort at the beginning of the test was also strongly associated with the initial perceived importance of the test. Additionally, change in effort was also highly related to change in the perceived value of the test during the testing session. In contrast to Barry and Finney (2016), who found no relationship between change in effort and change in importance during different cognitive and noncognitive tests, our study discovered evidence that changes in the TTM constructs were related to each other. Thus, *throughout a single cognitive test*, the different TTM constructs seem to be related, in contrast to TTM over different types of tests, which showed no relationship between change in one construct and change in another. Thus, from a theoretical and practical point of view it is important to assess all three components of EVT to capture the whole growth process of TTM. It appears that students show a smaller decline in effort than average if they also report a smaller decrease in perceived importance of the test throughout the testing session.

Change in TTM and Its Relationship to Test Performance

The last research question focused on the relationship between change in TTM and students' test performance, after controlling for students' backgrounds. Over 50% of the variance in mathematics performance was explained by students' background characteristics, with school track and grade in mathematics being the most predominant predictors. The final model added the growth parameters of probability of success, importance, and effort as well as self-concept in mathematics as predictor of test performance and the growth parameters of probability of success. In addition to reported effort before the test, the initial level of probability of success

and the change in probability of success, while controlling for self-concept in mathematics, also predicted test performance. Although the average change in probability of success over the testing session was fairly stable the interindividual variability of intraindividual change in probability of success was high enough that it revealed a relationship with test performance. Students who decrease less in their probability of success than the average tend to score higher in the cognitive test.

Moreover, it is well known that a domain-specific self-concept is related to performance in the corresponding domain (Chen, Yeh, Hwang, & Lin, 2013) and can affect probability of success (Asseburg, 2011; Eccles & Wigfield, 2002). Our study showed that beyond the stable domain-specific motivation, it is also important that students feel confident they will complete the test successfully. This result is consistent with the demands-capacity model of test-taking effort (Wise & Smith, 2011). The completion of previous test items can change the confidence in successfully completing the test, and in turn, the amount of effort students are willing to invest in further test items. Thus, a test booklet that includes alternating easy and difficult items throughout the test might be necessary to ensure a stable level of expectancy for success as well as high test performance among students.

Our study found an effect only for change in expectancy for success on test performance. However, change in importance of the test was an important predictor of change in effort. The proctor strategies presented by Lau and colleagues (2009) could be a promising means of increasing the perceived value of the test, and in turn, students' effort. The authors found that motivation enhancing behavior of the proctors (invigilators; such as emphasizing the importance and usefulness of the test; encouraging test-takers to give their full effort during the testing session) during the low-stakes testing session can affect invested effort. Emphasizing the importance and usefulness of the test aligns with emphasizing the value component of EVT. Our study provided support that attainment value and effort are strongly

related in that students who valued the test also showed higher effort and better test performance. We therefore recommend at least emphasizing the importance and usefulness of the test in the test instructions. In addition, the moderate change in TTM throughout the cognitive test indicated that a two-hour low-stakes test was not too exhausting for the students, although a longer test time might lead to a larger decrease in TTM. In this case, a further decrease in TTM could be due to the amount of fatigue or time pressure felt by students, which in turn might affect their willingness to invest effort on further items (Wise & Smith, 2011).

Limitations and Directions for Future Research

There are several limitations of our study that we address below. Although the strong measurement invariance of the TTM scales supported a successful adaption of the test items to different measurement points, more items per subscale would have been preferable (especially for the importance factor). Four items per subscale seems to be appropriate for second-order latent-growth modeling. Moreover, more measurement points are desirable, in order to test different growth forms in addition to linear change in TTM. For example, it is possible that a piecewise growth form fits the data in our study better, such as a larger decrease in TTM during the first half of the test and a smaller decrease in TTM during the second half of the test.

Furthermore, the measurement of the value scale can be optimized. We assessed indirectly the attainment value using the challenge scale. Most of the studies conducted internationally (Cole et al., 2008; Eklöf et al., 2013; Thelk et al., 2009; Wolf & Smith, 1995) assess the attainment value directly by asking students about their perceived importance of the test. It is conceivable that asking students directly how they perceive the test would lead to somewhat different results. Moreover, and as described above, the value component consists of four different aspects. In this study, only one aspect was included in the analyses. This is an

opportunity for future research. Furthermore, this study used self-report measures of TTM. As stated by Swerdzewski and colleagues (2011), such measures have several disadvantages: the test-takers a) need to recognize their current level of TTM, b) need to use the scale accurately to express their TTM, and (c) need to truthfully report their TTM. Despite these limitations, self-report measures are quite common in large-scale paper-and-pencil assessments.

Another limitation concerns the consideration of students' previous ability. Most previous research has found that the level of effort is not substantially related to high-stakes test scores when cognitive ability is controlled (Kong, Wise, Harmes, & Yang, 2006; Wise, Bhola, & Yang, 2006; Wise & Kong, 2005), but moderately related to low-stakes test scores (DeMars, Bashkov, & Socha, 2013). Our study used students' grade in mathematics as a measure of students' ability in mathematics prior to the test. We know that school grades account for not only intellectual capacity, but also for motivational and personality aspects; thus, grades are less objective than test scores on standardized achievement tests (Spinath, 2012). Ideally, we would like to control for prior knowledge with an additional measure from a high stakes test; however, this information was not available to us.

Furthermore, we did not assess TTM for the completion of the student questionnaire like previous studies (Barry & Finney, 2016; Barry et al., 2010; Horst, 2010). Instead, we used students' responses to draw conclusions about their attitude about school or to determine their socio-economic status. These data are generally very trustworthy and valid as they correspond strongly with the parents report about the socio-economic status (Jerrim & Micklewright, 2014). It is important that students also complete these questions with high effort. Further studies could compare TTM in large-scale assessments for both the cognitive test and the student questionnaire, as well as investigate change in TTM during the entire testing session, including the noncognitive test.

Conclusions

Our investigation of the change in TTM over the course of a cognitive large-scale assessment and its relationship to test performance based on EVT adds to the existing body of TTM research. We found an effect of initial TTM and, partly, an effect of change in TTM on test performance after taking into account students' socio-demographic background and their domain-specific motivation. Above all, it seems crucial that students begin the test with a high level of TTM and remain confident that they can complete the test successfully through the end of the testing session. To understand the mechanism of TTM during a testing session, it is important to assess all three components of EVT or one risks missing an essential TTM construct in low-stakes assessments.

Notes

[1] Due to the non-significant, slightly negative residual variances of some first-order factors in the second-order latent growth models, we had to fix some of the residual variances of the first-order factors to zero: for effort and importance for the first and third time point, and for probability of success for the third time point. An investigation of the residual variances using latent growth modeling with a composite of the manifest indicators per time point (instead of a latent variable) showed that these residual variances were close to zero. This supported our decision to fix the corresponding residual variances to zero.

[2] *Beta* refers to the stdyx standardization in the M*plus* output using full standardization with respect to both latent and observed variables.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Asseburg, R. (2011). *Motivation zur Testbearbeitung in adaptiven und nicht-adaptiven Leistungstests [Test-taking motivation in adaptive and sequential achievement testing]* (Doctoral dissertation). Christian-Albrechts-Universität zu Kiel. Retrieved from the website http://d-nb.info/1013153863/34

Barry, C. L., & Finney, S. J. (2016). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education, 29*(1)*, 46–64.* doi: 10.1080/08957347.2015.1102914

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*(4), 342–363. doi:10.1080/15305058.2010.508569

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*(3), 441–462.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Browne, M. W., & Du Toit, S. H. C. D. (1992). Automated fitting of nonstandard models. *Multivariate Behavioral Research*, *27*(2), 269–300. doi:10.1207/s15327906mbr2702_13

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*(2), 209–230. doi:10.1007/s11336-007-9045-9

Chen, S.-K., Yeh, Y.-C., Hwang, F.-M., & Lin, S. S. J. (2013). The relationship between academic self-concept and achievement: A multicohort–multioccasion study. *Learning and Individual Differences*, *23*, 172–178. doi:10.1016/j.lindif.2012.07.021

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. doi:10.1207/S15328007SEM0902_5

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*(4), 609–624.

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, *8*, 69–82.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. doi:10.1146/annurev.psych.53.100901.135153

Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement*, *66*(4), 643–656. doi:10.1177/0013164405278574

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, *7*(3), 311–326.

Eklöf, H. (2008). Test-taking motivation on low-stakes tests: A Swedish TIMSS 2003 example. In *Issues and methodologies in large-scale assessments, IERI Monograph Series* (Vol. 1, pp. 9–21). Hamburg: IEA-ETS Research Institute.

Eklöf, H. (2010a). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 345–356. doi:10.1080/0969594X.2010.516569

Eklöf, H. (2010b). *Student motivation and effort in the Swedish TIMSS Advanced field study*. Presented at the meeting of the 4th IEA International Research Conference, Gothenburg.

Eklöf, H., & Nyroos, M. (2013). Pupil perceptions of national tests in science: Perceived importance, invested effort, and test anxiety. *European Journal of Psychology of Education*, *28*(2), 497–510. doi:10.1007/s10212-012-0125-6

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2013). A cross-national comparison of reported effort and mathematics performance in TIMSS Advanced. *Applied Measurement in Education*, 131127082739006. doi:10.1080/08957347.2013.853070

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *4*(1), 22–36. doi:10.1027/1614-2241.4.1.22

Freund, P. A., & Holling, H. (2011). Who wants to take an intelligence test? Personality and achievement motivation in the context of ability testing. *Personality and Individual Differences*, *50*(5), 723–728. doi:10.1016/j.paid.2010.12.025

Freund, P. A., Kuhn, J. T., & Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personality and Individual Differences*, *51*(5), 629–634. doi:10.1016/j.paid.2011.05.033

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39–53. doi:10.1111/j.1745-3992.2009.00154.x

Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*(1), 1–56. doi:10.1016/0049-089X(92)90017-B

Geiser, C., Keller, B. T., & Lockhart, G. (2013). First- versus second-order latent growth curve models: Some insights from latent state-trait theory. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 479–503. doi:10.1080/10705511.2013.797832

Hancock, G. R., Kuo, W.-L., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 470–489. doi:10.1207/S15328007SEM0803_7

Horst, S. J. (2010). *A mixture-modeling approach to exploring test-taking motivation in large-scale low-stakes contexts* (Unpublished doctoral dissertation). James Madison University, Harrisonburg.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing*, *1*(2), 93–114. doi:10.1207/S15327574IJT0102_1

Jerrim, J. & Micklewright, J. (2014). Socio-economic Gradients in Children's Cognitive Skills: Are Cross-Country Comparisons Robust to Who Reports Family Background? European Sociological Review, *30*(6), 766–781. doi:10.1093/esr/jcu072

Kong, X. J., Wise, S. L., Harmes, J. C., & Yang, S. (2006). *Motivational effects of praise in response-time-based feedback: A follow-up study of the effort-monitoring CBT*.

Presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education*, *58*(3), 196–217. doi:10.1353/jge.0.0045

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, N.J: L. Erlbaum Associates.

Moneta, G. B., & Csikszentmihalyi, M. (1996). The effect of perceived challenges and skills on the quality of subjective experience. *Journal of Personality, 64*(2), 275–310. doi:10.1111/j.1467-6494.1996.tb00512.x

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide. Seventh edition*. Los Angeles, CA: Muthén & Muthén.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (2013). *The IQB National Assessment Study 2012. Competencies in mathematics and the sciences at the end of secondary level I. Summary*. Münster: Waxmann. Retrieved from http://www.iqb.hu-berlin.de/laendervergleich/laendervergleich/lv2012/Bericht/IQB_NationalAsse.pdf

Pekrun, R., Elliot, A. J., & Maier, M. A. (2009). Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance. *Journal of Educational Psychology*, *101*(1), 115–135. doi:10.1037/a0013383

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: an investigation of school-track-specific

differences. *Large-Scale Assessments in Education*, *2*(1). doi:10.1186/s40536-014-0005-4

Penk, C. & Schipolowski, S. (2015). Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Learning and Individual Differences, 42*, 27–35. doi:10.1016/j.lindif.2015.08.002Preacher, K. J., Wichmann, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Los Angeles: SAGE.

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., … Schiefele, U. (2006). *PISA 2003: Dokumentation der Erhebungsinstrumente [Documentation of the assessment instruments]*. Münster: Waxmann.

Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern-und Leistungssituationen [A questionnaire for the measurement of current achievement motivation in learning and achievement situations]. *Diagnostica*, *47*(2), 57–66.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. doi:10.1177/0013164413498257

Sayer, A. G., & Cumsille, P. E. (2001). Second-order latent growth models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (1st ed., pp. 179–200). Washington, DC: American Psychological Association.

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). *Motivation in education: Theory, research, and applications* (3rd ed.). Upper Saddle River, NJ: Pearson Education.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. doi:10.1007/s11336-008-9101-0

Spinath, B. (2012). Academic achievement. In V. S. Ramachandran (Ed.), *Encyclopedia of human behavior* (pp. 1–8). London; Burlington, MA: Elsevier/Academic Press.

Stanat, P., & Christensen, G. (2006). *Where immigrant students succeed: A comparative review of performance and engagement in PISA 2003*. Paris: Organisation for Economic Co-operation and Development.

Stanat, P., & Lüdtke, O. (2013). International large-scale assessment studies of student achievement. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 481–483). New York, NY: Routledge.

Sundre, D. L. (2007). The Student Opinion Scale: A measure of examinee motivation: Test manual. Retrieved from the Center for Assessment and Research Studies website: http://www.jmu.edu/assessment/resources/resource_files/sos_manual.pdf

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, *29*(1), 6–26. doi:10.1016/S0361-476X(02)00063-2

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*(2), 162–188. doi:10.1080/08957347.2011.555217

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. *The Journal of General Education*, *58*(3), 129–151. doi:10.1353/jge.0.0047

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-

concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, *98*(4), 788–806. doi:10.1037/0022-0663.98.4.788

Wainer, H. (2000). *Computerized adaptive testing: a primer* (2nd ed.). Mahwah, N.J: Lawrence Erlbaum Associates.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review*, *6*(1), 49–78. doi:10.1007/BF02209024

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. doi:10.1006/ceps.1999.1015

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114. doi:10.1207/s15324818ame1902_2

Wise, S. L., Bhola, D. S., & Yang, S. (2006). *Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT*. Presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. doi:10.1207/s15326977ea1001_1

Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. doi:10.1207/s15324818ame1802_2

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice.

Applied Measurement in Education, *22*(2), 185–205. doi:10.1080/08957340902754650

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: science and practice in K-12 settings* (1st ed., pp. 139–153). Washington, DC: American Psychological Association.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227–242. doi:10.1207/s15324818ame0803_3

Zilberberg, A., Finney, S. J., Marsh, K. R., & Anderson, R. D. (2014). The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: A path analytic model. *International Journal of Testing, 14*(4), 360–384. doi:10.1080/15305058.2014.928301

Table 1

Descriptive Statistics and the Reliability of the TTM Scales

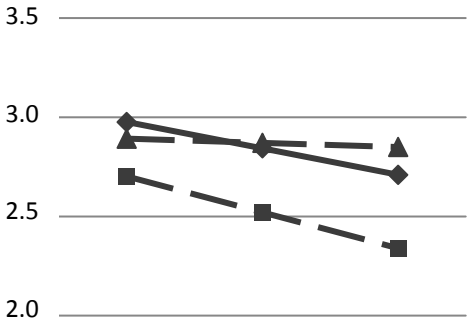
Scale		T1			T2			T3		
	N_{items}	*M*	*SD*	ω[1]	*M*	*SD*	ω[2]	*M*	*SD*	ω[3]
Effort	4	2.95	0.67	.83	2.73	0.75	.86	2.55	0.76	.85
Probability of success (E)	3	2.88	0.56	.64	2.85	0.66	.70	2.85	0.64	.67
Importance (V)	2	2.75	0.80	.67	2.47	0.86	.73	2.39	0.87	.75

Notes: *M* = mean; *SD* = standard deviation; ω = McDonald's Omega; E = expectancy for success; V = value; T1–T3 = measurement times before the test, half way through the test, and after the test, respectively. [1]N_{T1} = 42,080; [2]N_{T2} = 42,099; [3]N_{T3} = 22,601.

Table 2

Parameter Estimates and Model Fit for the Second-Order Latent Growth Curve Models for Effort, Importance, and Probability of Success

Factor		*M*	*SD*	Correlation between intercept & slope
Effort	Intercept	2.98*	0.47*	-.14*
	Slope (solid line)	-0.13*	0.16*	
Impor-tance	Intercept	2.70*	0.78*	-.06
	Slope (dotted line)	-0.18*	0.25*	
Probability of Success	Intercept	2.89*	0.41*	-.05
	Slope (dashed line)	-0.02*	0.18*	



	χ^2 (*df*)	CFI	TLI	RMSEA	SRMR	N
Effort	971.64* (54)	.99	.99	.02	.03	42,287
Importance	547.12* (7)	.98	.97	.04	.03	42,281
Probability of Success	415.88* (25)	.99	.98	.02	.04	42,292

Notes: *$p < .001$. *M* = mean; *SD* = standard deviation; T1–T3 = measurement times before the test, half way through the test, and after the test, respectively; *df* = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

Table 3

Correlations between the Intercept and Slope Factors of the Multivariate Second-Order Latent Growth Curve Model with Effort, Importance, and Probability of Success

Factor	Factor correlations					
	1.	2.	3.	4.	5.	6.
1. Effort intercept	–					
2. Effort slope	-.12*	–				
3. Importance intercept	.79*	-.04	–			
4. Importance slope	-.08*	.79*	-.04	–		
5. Probability of success intercept	.21*	-.04	.00	-.01	–	
6. Probability of success slope	.08*	.45*	.09*	.33*	.01	–

Notes: *$p < .001$.

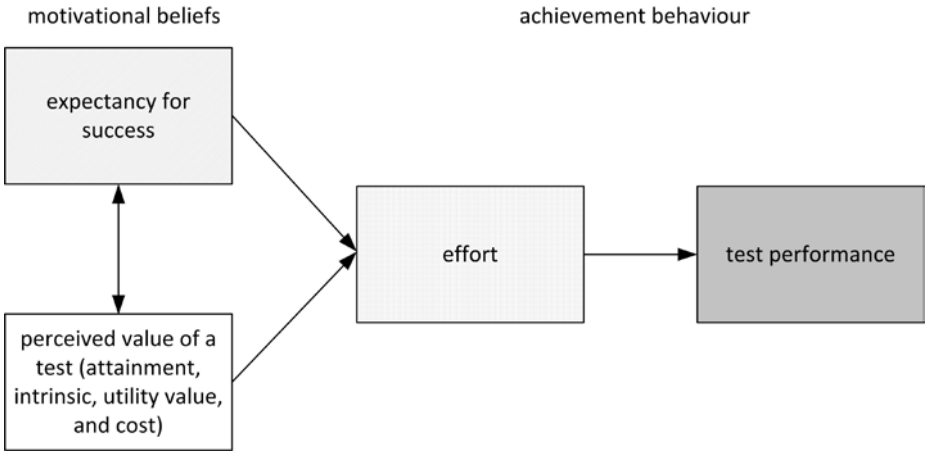


Figure 1. Expectancy-value theory in the context of test-taking motivation (adapted from Eccles & Wigfield, 2002; Wigfield & Eccles, 2000).

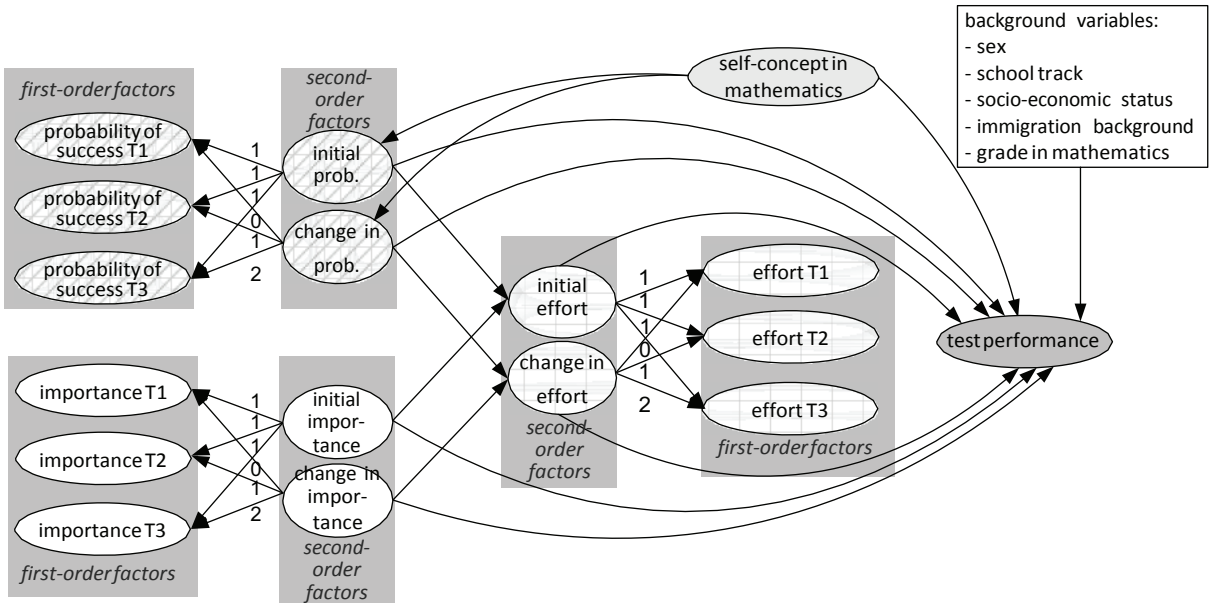


Figure 2. Prediction of test performance with the growth factors for probability of success, importance, and effort, controlling for students' background characteristics and self-concept in mathematics. Notes: Manifest indicators, disturbance terms, and correlations are omitted for simplicity. prob. = probability of success; T1–T3 = measurement times before the test, half way through the test, and after the test, respectively.

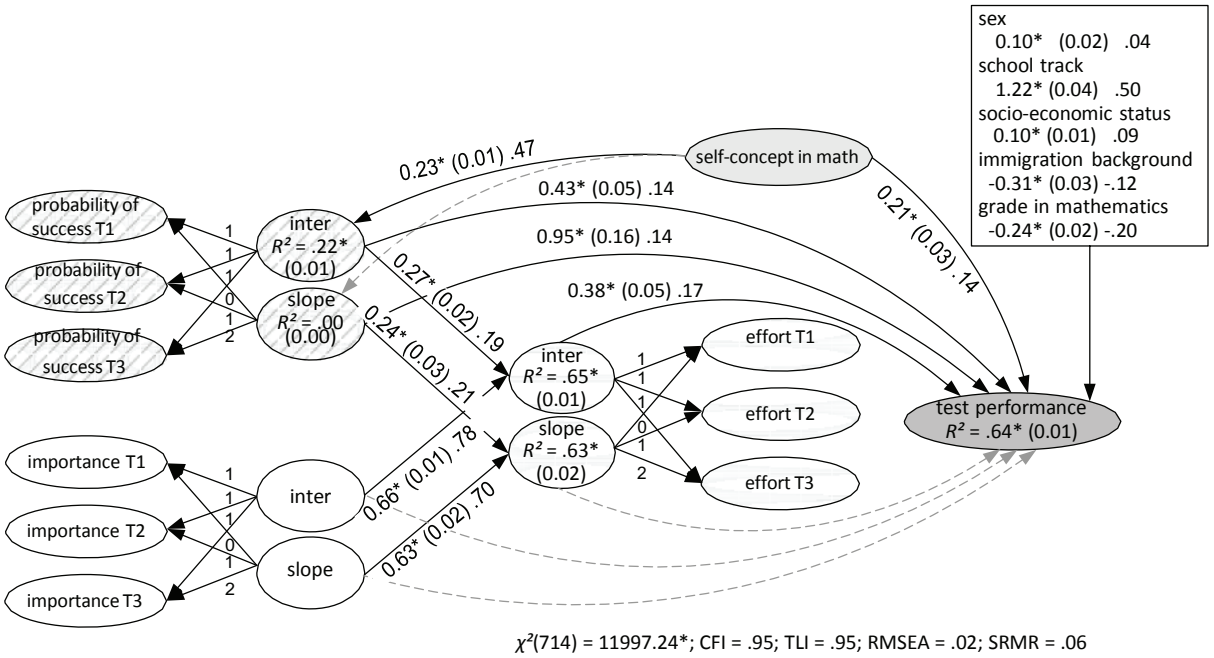


Figure 3. Prediction of test performance with the growth factors for probability of success, importance, and effort, controlling for the students' backgrounds and self-concepts in mathematics (Model 2). Notes: Manifest indicators, correlations, and the measurement and residual errors are omitted for simplicity. Coefficients: *unstandardized coefficient* with *p (standard error) standardized coefficient*. *$p < .001$. inter = intercept; T1–T3 = measurement times before the test, half way through the test, and after the test, respectively; T3 = after the test. Non-significant paths are dashed, but the coefficients are listed in Appendix B.

Appendix A

Table A1

Test of the Strong Measurement Invariance Test of Test-Taking Effort and Probability of Success, with Autocorrelated Errors

		χ^2 (*df*)	CFI	TLI	RMSEA	SRMR	RDR	*ΔCFI*
Effort	Configural	419.9* (39)	.996	.993	.015	.016	-	-
	Metric	563.1* (45)	.995	.992	.016	.023	.026	-.001
	Strong	743.2* (51)	.993	.992	.018	.025	.028	-.002
Expectancy for success	Configural	102.2* (15)	.997	.993	.012	.012	-	-
	Metric	156.9* (19)	.995	.991	.013	.019	.018	-.002
	Strong	265.9* (23)	.991	.986	.016	.025	.027	-.004

Notes: * $p < .001$. *df* = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; RDR = root deterioration per restriction statistic.

Appendix B

Table B1

Correlations of the Growth Parameters for Effort, Importance, Probability of Success, and Self-Concept in Mathematics: The Indirect Effects and Non-Significant Effects for Model 2

Correlations	*beta*	*(SE)*	
Effort intercept with effort slope	-.23*	(0.00)	
Probability of success intercept with probability of success slope	.03	(0.00)	
Importance intercept with importance slope	-.08*	(0.00)	
Importance intercept with probability of success intercept	-.02	(0.00)	
Importance slope with probability of success slope	.34*	(0.00)	
Effort intercept with self-concept in mathematics	.05	(0.01)	
Effort slope with self-concept in mathematics	-.03	(0.00)	
Importance intercept with self-concept in mathematics	.09*	(0.01)	
Importance slope with self-concept in mathematics	-.01	(0.00)	
Indirect effects	*b*	*(SE)*	*beta*
Performance on importance intercept via effort intercept	0.25*	(0.03)	.13
Performance on importance slope via effort slope	0.10	(0.13)	.02
Performance on probability of success intercept via effort intercept	0.10*	(0.01)	.03
Performance on probability of success slope via effort slope	0.04	(0.05)	.01
Non-significant effects	*b*	*(SE)*	*beta*
Performance on importance intercept	-0.09	(0.04)	-.05
Performance on importance slope	0.06	(0.18)	.01
Performance on effort slope	0.16	(0.20)	.03
Probability of success slope on self-concept in mathematics	0.00	(0.01)	-.02

Notes: *$p < .001$. *b* = unstandardized regression coefficient; *SE* = standard error; *beta* = standardized regression coefficient.

Table B2

Correlations of the Growth Parameters for Effort, Importance, Probability of Success, and Self-Concept in Mathematics with the Background Variables for Model 2

	Sex		School track		Migration background		Socio-economic status		Grade in mathematics	
	beta	*(SE)*	*beta*	*(SE)*	*beta*	*(SE)*	*beta*	*(SE)*	*beta*	*(SE)*
Effort intercept	.00	(0.01)	.14 *	(0.02)	-.08 *	(0.02)	.10 *	(0.01)	-0.05*	(0.01)
Effort slope	.02	(0.02)	.00	(0.02)	-.09 *	(0.02)	-.01	(0.02)	0.01	(0.02)
Importance intercept	-.14 *	(0.01)	.00	(0.02)	.03	(0.01)	-.04	(0.01)	-0.08*	(0.01)
Importance slope	-.07 *	(0.01)	.09 *	(0.01)	-.03	(0.02)	.06 *	(0.01)	-0.04	(0.01)
Probability of success intercept	.16 *	(0.01)	.16 *	(0.02)	-.11 *	(0.01)	.11 *	(0.01)	0.05*	(0.01)
Probability of success slope	-.05	(0.02)	.34 *	(0.02)	-.06 *	(0.02)	.16 *	(0.02)	0.04	(0.02)
Self-concept in mathematics	.26 *	(0.01)	.05 *	(0.01)	-.02	(0.01)	.06 *	(0.01)	-0.59*	(0.01)

Notes: *$p < .001$.*beta* = standardized regression coefficient; *SE* = standard error.