Using large-scale educational data to test motivation theories: a synthesis of findings from Swedish studies on test-taking motivation

Hanna Eklöf* and Eva Knekta

Department of Applied Educational Science,
Umeå University,
Umeå 901 87, Sweden
Email: hanna.eklof@umu.se
Email: eva.knekta@umu.se
*Corresponding author

Abstract: With the purpose of mapping the test-taking motivation construct, theoretically and empirically, the current paper presents a synthesis of findings from test-taking motivation studies performed in various large-scale educational assessment settings. Self-report scales measuring reported motivation to spend effort on the test, and in some studies also other aspects of expectancy and value, have been administered to Swedish samples taking tests such as the Trends in International Mathematics and Science Study – Advanced (TIMSS), the Programme for International Student Assessment (PISA) and the Swedish National Tests. Taken together, results indicate that tests with lower stakes are associated with a lower level of reported test-taking motivation, that the expectancy-value theory seems a viable framework for interpreting the test-taking motivation construct, and that reported level of motivation consistently seems to have a significant effect on test performance, also when modelled together with other motivational and background variables.

Keywords: effort; expectancy-value theory; national tests; PISA; research synthesis; structural equation modelling; test performance; test-taking motivation, TIMSS.

Reference to this paper should be made as follows: Eklöf, H. and Knekta, E. (2017) 'Using large-scale educational data to test motivation theories: a synthesis of findings from Swedish studies on test-taking motivation', *Int. J. Quantitative Research in Education*, Vol. 4, Nos. 1/2, pp.52–71.

Biographical notes: Hanna Eklöf is an Associate Professor at the Department of Applied Educational Science, Unit for Measurement in the Behavioural Sciences at Umeå University, Sweden. She has an MA in psychology and a PhD in Educational Measurement. Her research interests include the psychology of test-taking, international comparative studies (TIMSS, PISA) and measurement quality issues. She currently is the Leader of a research project funded by the Swedish Research Council, focusing on test-takers' perceptions of large-scale educational tests, test-taking effort and test-taking strategies and the relationship between non-cognitive variables and test performance.

Eva Knekta is a PhD student at the Department of Applied Educational Science, Unit for Measurement in the Behavioural Sciences at Umeå University, Sweden. She has a Licentiate Degree of Science in Chemistry and a Master of Science in Education. Her research subject is educational measurement, and she works foremost with students' attitudes towards

achievement tests. Her research interest includes students' perceptions of tests and test validity, and her current research project has aimed to investigate students' test-taking motivation at tests with different stakes, with a particular focus on the Swedish national tests.

This paper is a revised and expanded version of a paper entitled 'Different stakes, different motivation? Studies of test-taking motivation in different assessment contexts', presented at *the annual meeting of the American Educational Research Association (AERA)*, April 2014, Philadelphia, PA.

1 Introduction

How are different large-scale educational achievement tests perceived by test-takers? Are test-takers in general motivated to do their best, regardless of the characteristics of the test they are taking? How can the test-taking motivation construct be theoretically understood and modelled? And how is test-taking motivation related to test performance and thereby to score validity in different assessment contexts? Questions like these have served as the impetus to a number of empirical studies done in different Swedish large-scale assessment contexts over the last decade. In the following, findings from these different studies will be summarised, synthesised and discussed. The rationale behind this synthesis of findings is an ambition to explore the test-taking motivation construct more thoroughly by comparing and linking studies with different aims and performed in different contexts to each other, and thereby provide 'a bigger picture' of test-taking motivation: from its conceptualisation to how it relates to other variables in the assessment situation. Ultimately, the aim is to contribute to the understanding of the construct and how it operates in different assessment contexts. Most of the results presented in the following are based on unpublished data, while some of the results draw on findings that have been presented in other contexts and in other forms as part of other research questions. Given the aim and the character of the study, also these findings were, however, deemed important to include in the below presentation and discussion.

Motivation is an important driving force behind an individual's learning and achievement in different contexts, and motivated individuals tend to be more goal directed and persistent than less motivated individuals (Wentzel and Wigfield, 2009; Wigfield and Eccles, 2000). Motivation can be defined in different ways and measured at different levels. Test-taking motivation is conceptualised and measured at a situation-specific level and can be defined as "the willingness to work on test items and to invest effort and persistence in this undertaking" (Baumert and Demmrich, 2001, p.441). Hence, test-taking effort is assumed to be the outcome of a motivational state occurring in a specific assessment situation. This state in turn is assumed to be distinct from more domain-specific trait-like motivation, and to be related to performance in the test situation (Penk and Schipolowski, 2015). Another assumption is that test-taking motivation is particularly important to consider in test situations where the outcome is without consequences for the test-taker (low-stakes tests). When the test result has no consequences, some test-takers might not be motivated to do their best, and the result might be that students perform below their actual ability. Not taking test-taking motivation into account in these test situations can thus lead to an incorrect interpretation of the results, as test-taking motivation may have a construct-irrelevant impact on test

performance and therefore affect the reliability and validity of test scores (Attali, 2016; Finn, 2015).

As noted by Wise and Smith (2016), the importance of test-taking motivation has been recognised for more than a century as evidenced in writings by scholars such as Thorndike (in the early 19th century) and Lee Cronbach (in the middle of the same century), as well as by organisations such as American Educational Research Association (AERA), American Psychological Association (APA) and National Council for Measurement in Education (NCME) (2014). However, only during the last couple of decades has there been a notable body of research on this construct. Many studies concerning test-taking motivation have used the expectancy-value theory of achievement motivation as an interpretative/explanatory framework (Barry et al., 2010; Cole et al., 2008; Finn, 2015; Hawthorne et al., 2015; Knekta and Eklöf, 2015; Penk et al., 2014; Smith and Smith, 2002; Sundre and Kitsantas, 2004). The expectancy-value theory of achievement motivation is a theory designed to explain motivational determinants of achievement-related behaviour (Eccles et al., 1983). According to the theory, the most direct influences on students' performance, effort, persistence and task choice are the students' expectations for success and the value placed on the task (Wigfield and Eccles, 2000). Expectancies and task value are themselves influenced by students' goal and self-schemas as well as their affective memories. These social cognitive variables, in turn, are influenced by aspects such as the child's perception of socialisers' beliefs, gender roles and child's interpretations of experience. Applied on a specific test, the individuals' beliefs about how well they will do on the test (expectancies) and the reasons a student have to engage in the test (task value) will affect the effort the student invest in a test. Understanding both expectancies and values is critical for predicting students' willingness to engage in a test. A student might find a test interesting or valued for other reasons but if she does not expect to do well she might not engage in the task. Likewise, a student might be confident that she can do well on a test but if she does not value the task she still might chose not to engage with it. Students can have several reasons to why they want to engage in a test. According to Eccles et al. (1983), task value is influenced by four main components: importance (personal importance of doing well), interest (the inherent, immediate enjoyment one gets from engaging in in the test), utility (importance of the test for some future goal, such as get entry to a course) and cost (the negative aspects of engaging in the test including negative emotions, loss of valuable time and negative emotions).

Research on test-taking motivation based on the expectancy value theory have in concordance with the theory showed that students report higher effort on test that have consequences for them compared to test with no consequences (Immekus and McGee, 2016; Knekta, 2017; Liu et al., 2012, 2015; Thelk et al., 2009; Smith and Smith, 2002; Wolf and Smith, 1995). 'Consequences' in these studies usually means that the test score counts towards student marks. Research has also shown differences in performance depending on whether a test is performed under low- or high-stakes settings (Abdelfattah, 2010; Duckworth et al., 2011; Rutkowski and Wild, 2015; Wise and DeMars, 2005). A previous review of studies on test-taking effort and motivation suggested that test-taking motivation (high vs. low) makes an average difference in test score corresponding to about 0.5 SD (Wise and DeMars, 2005), a substantial difference in many assessment contexts. Other studies, however, failed to demonstrate clear relationships between test stakes, test-taking motivation and test performance (Baumert and Demmrich, 2001; O'Neil et al., 2005). Students test-taking motivation also seems to vary within low-stakes

test situations depending on both characteristics of the test and the test-takers. Constructed-response items and essay format have been shown to be associated with lower levels of motivation and effort in comparison to multiple-choice items (DeMars, 2000). Effort spent during a test has also shown to differ between females and males, students in different school classes and ethnic groups (Chan et al., 1997; DeMars et al., 2013; Knekta, 2017).

Thus, as evidenced by the above-reviewed research, several empirical studies have supported the assumptions surrounding test-taking motivation, but as motivation and attitudes towards schoolwork can be assumed to vary over age, subject and cultures (Dotterer et al., 2009; Wigfield et al., 2009), the test-taking motivation research field would still benefit from further empirical investigations in ecologically valid and varied large-scale achievement test situations. Also, although most research concerning test-taking motivation refers to the expectancy-value theory, most studies have only measured task value on a general level (e.g. 'this was an important test to me') and effort invested. Few studies have tested the effect of all the different task value components and expectancies on effort and thus whether the theory provides a suitable model for understanding test-taking motivation in a wider perspective and for predicting effort (see however Cole et al., 2008; Penk and Schipolowski, 2015). Furthermore, test-taking motivation in international, large-scale studies such as Trends in International Mathematics and Science Study (TIMSS) and Programme for International Student Assessment (PISA) is not very well researched despite the potential importance of the issue in these contexts (Butler and Adams, 2008; Eklöf, 2007; Hopfenbeck and Kjaernsli, 2016), given the impact of these studies on educational policy and practice (Rutkowski and Rutkowski, 2016). In addition, with the exception of studies by Penk et al. (2014) and Eklöf (2007) who modelled test-taking motivation together with self-concept, very few previous studies have modelled test-taking motivation with other domain-specific motivational variables in large-scale educational contexts.

The research that will be presented in the following explores all of the above. The research has been organised around the following assumptions and in the results section, we present findings accordingly:

i Test-taking motivation is lower on low-stakes test compared to high-stakes test.

ii The expectancy-value theory provides a suitable model for understanding test-taking motivation and for predicting test-taking effort.

iii Reported test-taking effort is positively correlated with test performance over samples and assessment contexts.

iv Test-taking effort is distinct from other domain-specific motivational variables and with a unique contribution to performance.

2 Methods

2.1 Samples and instruments

As noted, the current paper presents a synthesis of performed studies. In all studies, self-report questionnaire data were collected together with achievement test data from large-scale assessments, i.e. TIMSS Advanced 2008 (field trial and main study), PISA 2012,

the Swedish Scholastic Assessment Test (SweSAT, nationally used for admission to higher education), Swedish National Tests in the science subjects (field trial and regular test) in lower secondary school and Swedish National Tests in mathematics (field trial and regular test) in upper secondary school. An overview of the data sources, the samples and the measures of test-taking motivation used in the respective data collections is presented in Table 1.

Table 1 Reviewed studies, including their stakes in terms of test score consequences for the individual test-taker, the sample composition and the measures of test-taking motivation used

Study	*Stakes*	*Sample*	*TTM measure*
SweSAT 2012	High	Median age: 20 years, n = 1,778	Effort (3 items)
			Importance (3 items)
			Test Anxiety (4 items)
National Test 2014 (Mathematics)	High	Grade 11 (17–18 years), n = 81	Effort (4 items)
			Importance (3 items)
			Expectancy (3 items)
			Interest (3 items)
			Utility (2 items)
			Loss of valuable time (2 items)
			Test anxiety (2 items)
National Test 2012 (Science)	High	Grade 9 (15–16 years), n = 536	Effort (4 items)
			Importance (3 items)
			Expectancy (2 items)
			Interest (3 items)
			Test Anxiety (3 items)
PISA 2012	Low	Grade 9 (15–16 years), n = 4,129	Effort (4 items)
			Importance (2 items)
National Test 2012 - field trial (Science)	Low	Grade 9 (15–16 years), n = 1,422	Effort (4 items)
			Importance (3 items)
			Expectancy (2 items)
			Interest (3 items)
			Test Anxiety (3 items)
National Test 2014 - field trial (Mathematics)	Low	Grade 11(17–18 years), n = 222	Effort (4 items)
			Importance (3 items)
			Expectancy (3 items)
			Interest (3 items)
			Utility (2 items)
			Loss of valuable time (2 items)
			Test anxiety (2 items)
TIMSS Advanced 2008 main study	Low	Grade 12 (18–19 years), n = 2,218	Effort (6 items)
TIMSS Advanced 2008 field trial	Low	Grade 12 (18–19 years), n = 163	Effort (6 items)
			Importance (3 items)
			Test Anxiety (3 items)

Although studies such as TIMSS and PISA can be viewed as high-stakes on a political level, they are low-stakes tests for the participating students. Students participating in TIMSS or PISA must devote hours to a test that for some might be quite demanding, while the result of this investment has no practical significance for them personally. The result does not count towards students' mark, and they receive no individual feedback, but results are reported only at the country level long after the study has been conducted. The purpose of the Swedish National Test field trials is to, through separate field trials, try out items that are subsequently going to be used in the regular national tests. Teachers are asked to administer the field trial under test-like conditions and to urge the students to give their best effort to the test. However, it is generally considered that the field trials are performed under conditions that are lower stakes compared to the regular test, which was also confirmed by Knekta (2017). The regular national tests, on the other hand, and the SweSAT, are regarded as high-stakes in terms of the consequences for the individual test-taker (national test results are taken by all students participating in regular lower secondary schooling/by all students in certain upper secondary school programmes, and they are supportive of student marks in the subject tested, SweSAT scores are used for admission to higher education).

As can be seen in Table 1, different studies have used different measures of test-taking motivation, depending on the purpose of the study and/or practical circumstances such as space available. In some studies, such as TIMSS and PISA, reports on test-taking effort only or effort and importance were collected. The scales for measuring effort and importance have drawn on items from the Student Opinion Scale, originally developed by Wolf and Smith (1995) and later refined and extensively validated and used by Sundre and Finney (2002) and Sundre and Moore (2002) as well as items previously used to study test-taking motivation in the TIMSS 2003 context (Eklöf, 2006). In the studies in the National Test contexts, more elaborate measures were used, enabling more thorough tests of the theoretical assumptions regarding test-taking effort from an expectancy-value perspective. Furthermore, in PISA 2012, although only a brief measure of test-taking effort was used, students complete a questionnaire with a range of domain-specific motivational variables, enabling the modelling of test-taking effort with other variables to an extent that has rarely been done in previous research. Thus, taken together, synthesising findings from these data collections may further the understanding of the test-taking motivation construct; the relationships between the different parts of the construct and the relationship with performance.

2.2 Procedure and data analysis

The test-taking motivation measures have been completed directly after the achievement test, either alone (the more elaborate measures in the national test contexts) or as part of other questionnaires (TIMSS, PISA, SweSAT). In the national test contexts, partly the same sample of students completed the same test-taking measure at the field trial (low-stakes condition) and the regular test (high-stakes), making it possible to compare ratings from the same individuals taking tests with different stakes. In the analyses below, a particular focus will also be on comparing findings from PISA 2012 to findings from the regular national test 2012. These studies were performed in the same student cohort (although not necessarily with the same individual students, and with no information available about the identities of students in PISA, why samples are treated as independent in the analyses): students who were in the 9th school year in 2012, taking tests with

different stakes and completing either a short scale assessing test-taking effort (in low-stakes PISA) or participated in a more conceptual study of test-taking motivation (in the higher-stakes National Test context).

Primary and secondary analyses of the data have been undertaken with IBM SPSS, IEA IDB analyser (TIMSS and PISA data) and MPlus 7.11 (Muthén and Muthén, 1998-2012). To explore whether test-takers taking low-stakes tests in general report lower levels of motivation than test-takers taking high-stakes tests, descriptive statistics such as percentage agreement on the item level, mean scale scores and tests of mean differences as well as effect sizes (d) have been computed. The analysis of the relationship between reported effort and test performance across the different test contexts is based on correlation analyses.

To test the theoretical assumptions of the expectancy-value theory and to model test-taking effort together with other variables as predictors of performance in PISA, confirmatory factor analyses (CFA) and structural equation modelling (SEM) have been employed. In the respective analyses (using data from National Tests 2012, National Tests 2014 and PISA 2012, respectively), maximum likelihood estimation with robust standard errors (MLR) was used. For each model, structure and pattern coefficients, communalities, residuals, and modification indices were examined and multiple fit indices (MLR $\chi 2$, comparative fit index, CFI; the root mean squared error of approximation, RMSEA; and the standardised root-mean squared residual, SRMR) were consulted to evaluate model fit. CFI > 0.90, SRMR < 0.10 and RMSEA < 0.08 was considered adequate and CFI > 0.95, SRMR < 0.05 and RMSEA < 0.05 was considered good fit (Schermelleh-Engel et al., 2003).

3 Results

The results section will be organised around the assumptions that have guided the research synthesised in this article and the overall findings following from exploring these assumptions:

i Test-taking motivation is lower on low-stakes tests.

ii The expectancy-value theory can be used to explain test-taking effort.

iii Test-taking effort is related to test performance.

iv Test-taking effort is distinct from domain-specific motivation.

Hence, the results section starts with a general description of findings as concerns reported motivation on tests with different stakes, administered on a large scale in Swedish samples. In this overview, findings are exemplified by responses to a number of items that have been used in all studies. After that, findings from two studies testing the expectancy-value assumption that student effort is "a function of perceived expectations of success and the perceived value of doing well on the test" (Wise and DeMars, 2005, p.15) will be briefly summarised, followed by a review of findings regarding the relationship between reported test-taking effort and test performance in different Swedish large-scale test contexts. Last, an analysis will be presented where situation-specific test-taking effort has been modelled together with other background and domain-specific motivational variables as predictors of performance in PISA 2012.

3.1 Students are less motivated to give their best effort when the stakes of the test are low

In general, when findings over different studies are summarised, they rather convincingly suggest that tests with higher stakes are associated with higher ratings of motivation. To illustrate this, percentage agreement on a number of selected items (*I was motivated to do my best on this test; I worked with all the items in the test without giving up, even when a task was difficult; It was important for me to do well on this test*) are displayed in bar charts below. In all studies but one (SweSAT 2012), students have rated motivation on a 4-point scale with labels 'agree a lot', 'agree a little', 'disagree a little', 'disagree a lot' and below, the two agree options have been collapsed into one agree category. In the SweSAT context, a 5-point scale was used, ranging from 'completely disagree' to 'completely agree', why the results are not directly comparable and must be interpreted with caution. The SweSAT data displayed in the figures are the proportion of students who either 'somewhat', 'to a large extent' or 'completely' agreed with the different statements.

Figure 1 Percentage agreement* with the statement '*I felt motivated to do my best on this test'* in five low-stakes and three high-stakes assessment contexts. *Agree a lot + agree a little for all studies but SweSAT where the agree proportion is based on 'completely agree + agree to a large extent + somewhat agree' (see online version for colours)

In all studies except TIMSS Advanced (students in their 12th school year, taking advanced courses in mathematics or physics) a majority of student agreed with the statement that they felt motivated to do their best on the test at hand. There is however a marked distinction between tests that can be regarded as low-stakes for the students (national test field trials, PISA) on one hand and high-stakes tests (regular national tests, SweSAT) on the other. For example, the difference (expressed in effect size *d*) in percentage agreement on this item for students participating in PISA and students from the same cohort taking a national test in science is $d = 0.8$, a large difference (Cohen, 1969).

The same pattern as revealed in Figure 1 is evident in Figures 2 and 3, and other items in the scales used in the different studies also display the same pattern. Students participating in TIMSS Advanced (field trial as well as main study) report lower levels of

motivation and effort compared to the other samples taking low-stakes tests, but all samples taking low-stakes tests report significantly lower levels of test-taking motivation than the samples taking high-stakes tests. Effect sizes between tests with different stakes, as exemplified by the comparisons between PISA 2012 and the 2012 regular national test, are large, in particular for perceived importance of the test.

Figure 2 Percentage agreement* with the statement '*I worked with all items in the test without giving up, even when a task was difficult'* in five low-stakes and three high-stakes assessment contexts. *Agree a lot + agree a little for all studies but SweSAT where the agree proportion is based on 'completely agree + agree to a large extent + somewhat agree' (see online version for colours)

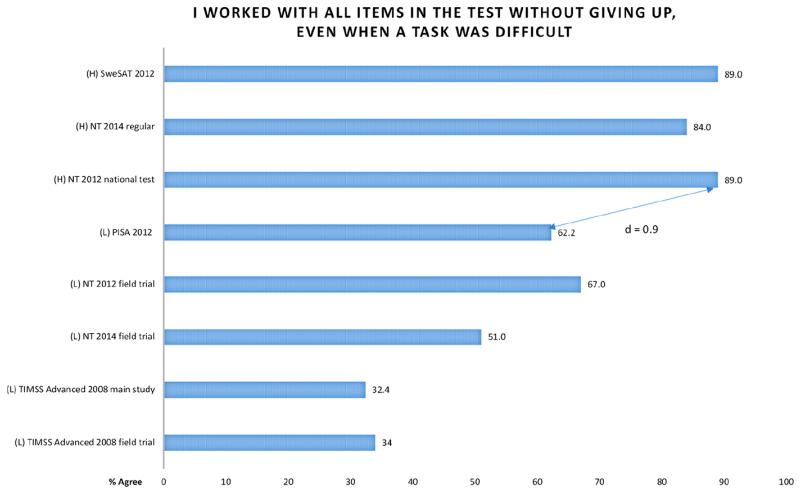
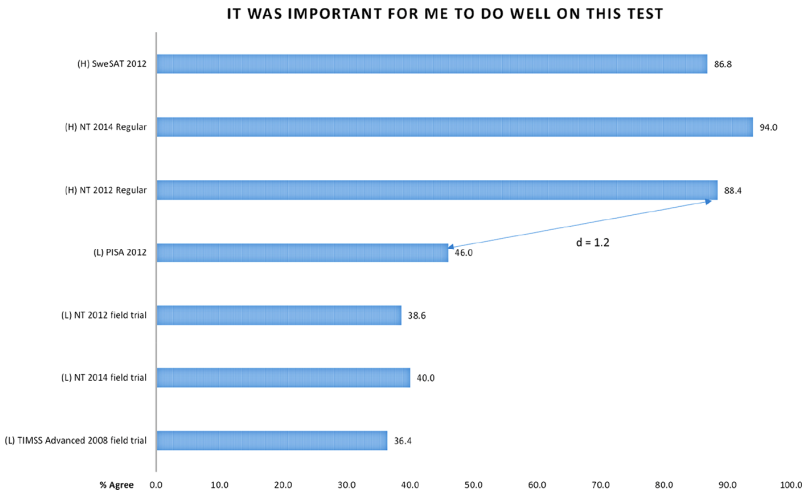


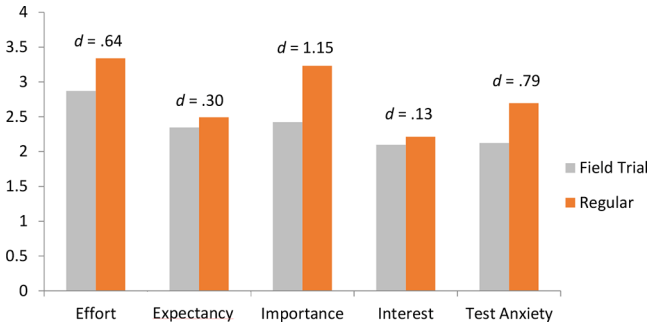
Figure 3 Percentage agreement* with the statement '*It was important for me to do well on this test'* in five low-stakes and three high-stakes assessment contexts. *Agree a lot + agree a little for all studies but SweSAT where the agree proportion is based on 'completely agree + agree to a large extent + somewhat agree' (see online version for colours)



As concerns TIMSS Advanced, cross-national comparisons have further shown that Swedish students participating in TIMSS Advanced 2008 tended to report lower levels of reported effort than students from Norway and Slovenia (Eklöf et al., 2014).

When calculating mean scale scores for the scales used to assess test-taking motivation (see Table 1 for an overview of the scales used in the different studies), there are also significant differences between scales that would be theoretically assumed to differ between low-stakes and high-stakes settings. For example, when comparing the same sample of students (n = 375) taking a national test field trial and a regular national test in science, respectively, effect sizes for differences in mean scale scores are moderate to be large for reported effort, perceived importance of the test and test anxiety, while they are small for interest and expectancy (see Figure 4 for details on psychometric properties of these scales; see Knekta and Eklöf, 2015).

Figure 4 Mean scale scores and difference effect sizes (d) for the same sample of Year 9 students (n = 375) reporting effort, performance expectancy, perceived importance, interest and test anxiety on a low-stakes (field trial) and a high-stakes (regular national) test, respectively (see online version for colours)

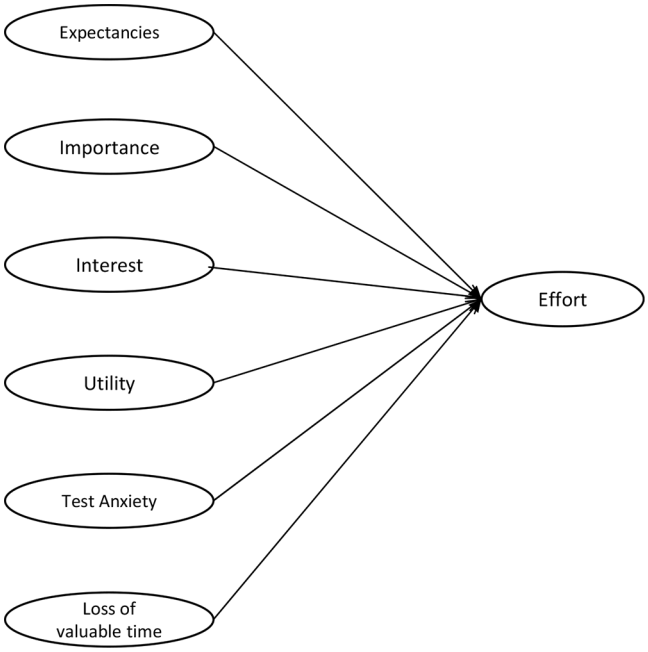


3.2 The test-taking motivation construct can be understood through expectancy-value theory, and test-taking effort is predicted by expectancies and values

Below, a brief summary of findings from studies will be presented, where the test-taking motivation construct has been modelled according to the propositions of the expectancy-value theory, with expectancies and value perceptions as predictors of 'motivated behaviour', in the test situation, i.e. effort. In this section, national test data only are discussed, as they contained more elaborate measures of test-taking motivation (see Table 1, Figure 5), designed with the ambition to model expectancies and a complete range of value aspects as predictors of test-taking effort. To test the fit of the data to the theoretical model in the Year 9, 2012 science national test field trial context, a five-factor model (with expectancies, importance, interest and test anxiety as predictors of effort) was specified and tested by means of SEM. The model showed an adequate fit to data (e.g. MLR χ^2 (80, n = 1047) = 432, $p < 0.001$, CFI = 0.92, RMSEA = 0.065, SRMR = 0.057). Also, reported effort was to a high degree predicted by the other four constructs: R^2 = 0.73 (for more detailed results from these analyses, see Knekta and Eklöf, 2015). The test-taking motivation instrument was then further elaborated on in accordance with the expectancy-value theory. For example, an additional cost scale, 'loss of valuable

time', was added, as was a utility scale (Figure 5). The instrument was administered to an upper secondary school sample (Year 11), participating in a field trial to the national test in mathematics and a seven-factor model (with expectancies, importance, interest, utility, loss of valuable time and test anxiety as predictors of effort) was specified and tested (Figure 5). This model also showed adequate fit to data (e.g. MLR χ^2 (131, n = 222) = 218, $p < 0.001$, CFI = 0.95, RMSEA = 0.055, SRMR = 0.061, and effort was to a rather high degree predicted by the other six subscales $R^2 = 0.52$.

Figure 5 Expectancy-value based theoretical model of the test-taking motivation construct with expectancies and a full set of value aspects as predictors of test-taking effort



Here, a somewhat counterintuitive finding was that the more elaborate test-taking motivation model actually explained somewhat less of the variation in the effort variable (R^2 = 0.52) compared to the less elaborate model (R^2 = 0.73). This could possibly be explained by the fact that the measures were used in different populations (Students in Year 9 taking science tests vs students in Year 11 taking a mathematics test), or by the fact that some of the items were revised from the first measure to the second.

Still, taken together, these findings support the assumption that the expectancy-value theory is a suitable framework for understanding the test-taking motivation construct, and that test-taking effort can be interpreted as the outcome of a motivated state, predicted by other aspects of motivation (although there remains variation in the effort scale that could not be explained by the other expectancy and value scales). Bearing this in mind, we now first turn to a brief summary of findings as regard the relationship between test-taking effort and test performance, and then to modelling test-taking effort in the PISA 2012 context.

3.3 Test-taking effort is associated with test performance

In all samples and all assessments included, low-stakes as well as high-stakes, there have been positive and significant correlations between test-taking motivation, expressed as test-taking effort and test performance (correlations in the range between $r = 0.16$ and 0.48 for the effort scale, see Table 2). The average correlation across the seven reviewed data collections in terms of correlation was $r = 0.30$.

Table 2 Correlations between reported test-taking effort and test performance across studies with different samples taking tests with different stakes

Study	*Stakes*	*Sample*	*Effort - test score r*
SweSAT 2012	High	Median age: 20 years	Effort (3 items) $r = 0.25$
National Test 2014 regular	High	Grade 11 (17–18 years)	Effort (4 items) $r = 0.32$
National Test 2012 regular	High	Grade 9 (15–16 years)	Effort (4 items) $r = 0.48$
PISA 2012	Low	Grade 9 (15–16 years)	Effort (4 items) $r = 0.27$
National Test 2012 field trial	Low	Grade 9 (15–16 years)	Effort (4 items) $r = 0.35$
National Test 2014 field trial	Low	Grade 11 (17–18 years)	Effort (4 items) $r = 0.16$
TIMSS Advanced 2008 main study	Low	Grade 12 (18–19 years)	Effort (6 items) $r = 0.30$

In a majority of the data collections drawn upon in the present paper, scales for assessing perceived importance of the test and test anxiety have also been used. Although importance also shows a positive correlation with performance in all studies where it has been assessed, this relationship has, within studies, generally been weaker than the relationship between effort and performance ($r = 0.07$–0.29). As concerns reported test anxiety, there has been a negative but often rather weak relationship between test anxiety and test performance ($r = -0.05$ to -0.26 in different studies) and in multiple regression analyses, the effect of test anxiety on performance has been smaller (and typically not significant) than the effect of effort. The exception was the SweSAT (a high-stakes test), where the (negative) effect of reported test anxiety on performance was stronger than the (positive) effect of reported effort on performance ($\beta = -0.29$ vs $\beta = 0.23$) (only test-taking motivation variables included in the regression model).

3.4 Situation-specific test-taking effort is distinct from domain-specific motivation and with a unique contribution to performance

Above, the test-taking motivation construct was modelled according to the propositions of the expectancy-value theory, and according to the assumptions of much previous research (Wise and DeMars, 2005; Finn, 2015), with expectancies and value perceptions as predictors of effort. These expectancy-value based propositions seem to hold in the tested assessment situations: a large proportion of reported effort could be explained by expectancy and different aspects of value, and students tend to report higher levels of test-taking motivation when the stakes of the test are higher. Correlation analyses also suggest that reported test-taking effort consistently has had a positive relationship with performance. Moving on from the 'within-construct' perspective in understanding test-taking motivation and test-taking effort, it seems relevant also to explore how the outcome of motivated behaviour, test-taking effort, behaves when it is modelled together

with other domain-specific motivational variables. As test-taking effort was assessed with four items in the Swedish PISA 2012 study, and PISA otherwise collects data on numerous background variables through questionnaires (see OECD, 2013 for a description of items and variables), some of which can also be interpreted from an expectancy-value perspective, this could be done in the context of PISA. Accordingly, the following domain-specific latent factors were included in a SEM model as predictors of mathematics performance in PISA (as mathematics was the main subject tested in 2012), together with the test-specific four-item Effort scale:

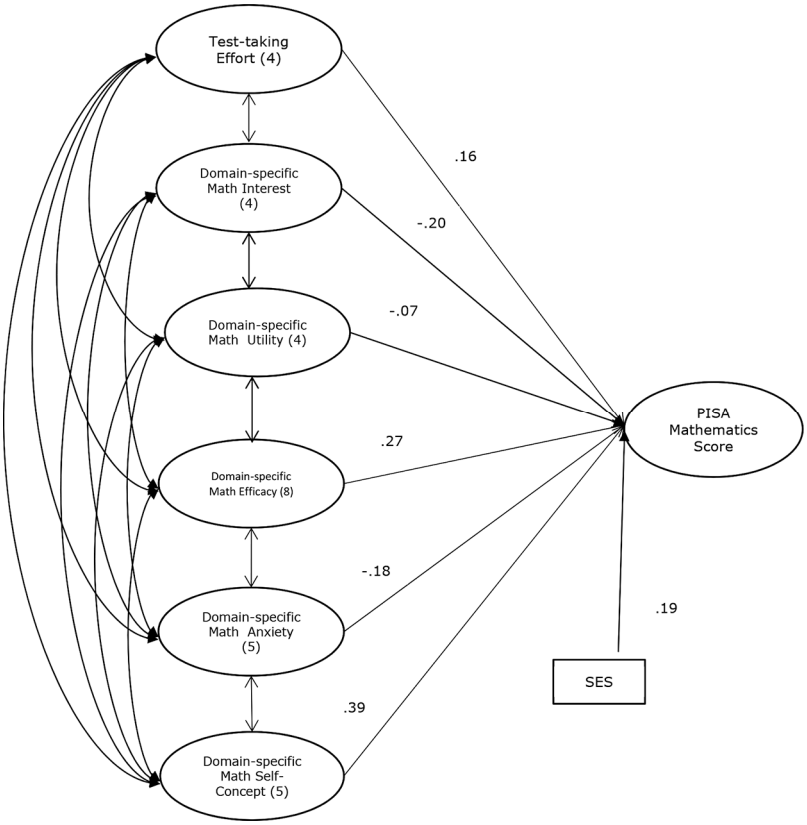
- *Mathematics self-concept:* whether students think they are good at mathematics and learn easily.
- *Mathematics Self-efficacy:* whether students feel confident they can solve mathematics tasks.
- *Intrinsic value:* interest in mathematics.
- *Utility value:* the value of mathematics for future studies/career.
- *Mathematics anxiety:* worries and other negative emotions associated with mathematics.

The analysis also included the PISA index of economic, social and cultural status (ESCS) to take into account socio-economic background. All these variables had significant correlations with mathematics performance when analysed separately (bivariate correlations, $p < 0.001$). Correlations were positive for all variables except math anxiety.

A model with the above variables (socio-economic background, five domain-specific motivation variables and one test-specific test-taking effort variable) was specified and tested in Mplus (Figure 6). Weighted data were used and all five plausible values were included as dependent variables (see OECD, 2014 for a description of the rationale behind using weights and plausible values as estimates of achievement scores). All independent latent variables were allowed to correlate. According to model fit indices, the model fitted the data well: χ^2 (573) = 2.274, $p < 0.01$, CFI = 0.95, RMSEA = 0.04, SRMR = 0.05. Reported test-taking effort was weakly to moderately correlated with the other domain-specific latent variables (ranging from $r = -0.14$ for math anxiety to $r = 0.30$ for math efficacy). Although there were rather strong intercorrelations between some of the other variables, for example math interest and self-concept ($r = 0.74$) interest and self-efficacy ($r = 0.55$) and interest and perceived utility of math ($r = 0.70$), respectively. As can be seen in Figure 6, mathematics interest has a negative effect on mathematics performance when modelled together with the other variables although it was positively related to performance when bivariate correlations were computed. That mathematics interest (one aspect of value according to expectancy-value theory) had strong correlations with a number of other variables in the model could explain this somewhat counterintuitive finding. Thus, it seems as if the effects of interest were suppressed by the strong effects of mathematics self-concept and mathematics self-efficacy. The same can be said to hold for the utility of mathematics factor that was also positively related to performance according to bivariate correlations, but slightly negatively related to performance when self-concept and self-efficacy were entered into the model. Test-taking effort, on the other hand, was a positive and significant predictor of mathematics performance in PISA, regardless of which other variables were entered into the model. The unstandardised regression coefficient for effort was $B = 20.22$,

implying that one points change on the (four-point) test-taking effort scale corresponds to a 20-point increase in mathematics score in PISA. The standardised effect of reported test-taking effort on test score was in the same range as that of socio-economic background and math anxiety (Figure 6).

Figure 6 Reported situation-specific test-taking effort modelled together with domain-specific motivational variables measured in the PISA student background questionnaire and socio-economic index as predictors of mathematics performance in PISA 2012. Number of items in the scales in parentheses, standardised pattern coefficients displayed in the figure



5 Summary and discussion

By synthesising a number of Swedish studies on test-taking motivation, performed in different large-scale achievement test contexts, the current paper has attempted to provide an overall picture of the test-taking motivation construct, from its conceptualisation to its relationship with other variables and with performance. Accordingly, the paper has summarised general findings across studies when it comes to reported level of test-taking motivation on tests with different stakes and on the relationship between reported level of test-taking effort and test performance. In addition, models testing theoretical propositions with expectancy and value aspects predicting test-taking effort were linked

to models where test-taking effort was related to domain-specific motivation in PISA and to PISA mathematics performance.

Results consistently show that students report lower levels of test-taking motivation on tests with lower stakes but also that in most samples, a majority of students reports spending effort also on low-stakes tests. This is a reassuring finding suggesting that many students do not simply dismiss low-stakes tests in terms of their willingness to invest effort. Similar findings have been reported in other studies (Hopfenbeck and Kjaernsli, 2016) and as noted by, for example, Wise and colleagues (Wise and Cotton, 2009; Wise and Smith, 2016), most students always give their best effort on all tests, possibly due to a sense of 'academic citizenship' (Kornhauser et al., 2014). Still, there seems to be a proportion of test-takers for whom the stakes of the test matter, and this proportion is important to acknowledge from a test-psychology perspective and a test validity perspective. When ratings from low- vs high-stakes tests were compared for students from the same cohort (PISA and national test samples), and sometimes also the same sample (taking a national test field trial and a regular national test), mean difference effect sizes were large for reported effort, perceived importance of the test and test anxiety, with d ranging from 0.6 to 1.2.

Of all samples, the lowest level of effort and motivation was reported by students participating in TIMSS Advanced 2008. The difference between TIMSS Advanced and the other low-stakes tests suggests that ratings of test-taking motivation on low-stakes tests may be related to age and experience with testing. PISA is 'as low-stakes' as TIMSS Advanced, but the Grade 12 students participating in TIMSS Advanced report lower levels of motivation than Grade 9 students taking the PISA test. It should, however, be noted that students in TIMSS Advanced represent a selected group of students, taking advanced courses in mathematics. They are in their last semester of secondary schooling, and presumably more tactical in their studies and aware of where they need to invest effort and not. They also differ in other ways from the group of students participating in PISA, why it may be difficult to make direct comparisons.

In all reviewed studies, test-taking motivation (expressed in terms of reported effort) has been significantly correlated with performance. Across studies, the average correlation between test-taking effort and performance was $r = 0.30$, a clear but modest relationship, and in the same range as reported in a previous synthesis of studies by Wise and DeMars (2005). Whether this significant but rather small effect is detrimental to the validity of the interpretation and use of test score may depend on the tolerance for error in the particular test at hand (Kane, 2011). In some assessment contexts, it might not matter that much if students perform a bit worse than they could have due to low motivation, while in other contexts, scores that are systematically biased downwards might have large consequences for the validity of score interpretation, even if test-taking motivation would make only a few points difference. Furthermore, the correlation between test-taking effort and performance presented in this article is averaged over individuals and different item types. As test-taking motivation has been shown to vary depending on personality and to have different effect on different types of items (Barry et al., 2010; DeMars, 2000), interpretations concerning item characteristics and student ability will most likely be more biased for some items/students than for others.

As noted in the introduction, test-taking motivation is often interpreted within an expectancy-value framework, and it is assumed that test-taking motivation is a special type of motivation, distinct from domain-specific motivation, even if these assumptions have more seldom been empirically explored. The present paper presents evidence that

test-taking effort to a rather high degree is predicted by expectancies and a range of value components, as modelled in the Swedish national test context, and that test-taking effort is distinct from domain-specific expectancies and values, as modelled in the PISA 2012 context. When test-taking effort was modelled together with other background and domain-specific motivational variables as predictors of performance in PISA, it was further shown that test-taking effort was a significant predictor of performance in PISA with an effect size in the same range as socio-economic background and math anxiety. Thus, test-taking motivation is one of many variables, above actual ability, that can affect test results.

Many large-scale educational assessments are low-stakes for the test-takers, but high-stakes for other stakeholders and in these contexts, valid interpretation of findings are important. The currently investigated contexts are Swedish, but the issues touched upon are general and relevant in a large range of assessment settings. For example, results from international comparative studies such as TIMSS and PISA are of great concern to policymakers all around the world (Rutkowski and Rutkowski, 2016). They are used to evaluate the quality of the educational systems, for cross-national comparisons and sometimes even as a ground for educational reform. Results from field trials/item try-outs are often used to evaluate items for the assembly of high-stakes tests, and it may be detrimental for the high-stakes test if results from the item try-outs are biased due to low motivation among test-takers. Also, in many countries, tests that are low-stakes for test-takers are increasingly used for accountability purposes or to make other important decisions at teacher, school or policy level (cf. Liu et al., 2015; Hawthorne et al, 2015; Sessoms and Finney, 2015). Given the findings presented in the current paper, test-taking motivation is clearly an issue that may be important to consider in these and other large-scale educational assessment contexts.

5.1 Limitations and future studies

The research presented in the current paper has some notable limitations, one being that the assessment of motivation is purely based on self-report with all the inherent limitations of this method (Wise and Cotton, 2009). More objective measures of test-taking effort, motivation and other behaviour, could further contribute to the understanding of the dynamics of test-taking psychology in large-scale assessment settings. Student motivation can vary over the course of a test and over different types of items (Wise and Smith, 2016), and this is difficult to capture via post-test self-report. As more tests are becoming computer-based, this gives opportunities to collect more objective response data together with self-report. For example, PISA is from the 2015 cycle computer-based, which opens up possibilities of analysing log files, response patterns and other test-taking behaviour on the item level. It should also be noted that although the theoretical models tested in the current paper provided a good fit to data, this does not mean that other models with other conceptualisations of the relationships between constructs could not also be plausible, and continued work on the theoretical underpinnings of motivation in the test situation seems important in order to better understand the dynamics of test-taking motivation.

Finally, if acknowledging that test-taking motivation is an important construct to consider in many large-scale assessment contexts, it follows that efforts should be taken to:

a raise student motivation

b in some way take test-taking motivation into account when interpreting findings.

Over the last decade, a number of methods for increasing motivation (Lau et al., 2009; Liu et al., 2015; Sessoms and Finney, 2015; Zilberberg et al., 2009) or statistically accounting or adjusting for 'non-effortful behaviour' (Sundre and Wise, 2003; Swerdzewski et al., 2011; Wise and Kong, 2005; Wise et al., 2009; DeMars, 2007) have been explored. Some of the interventions and statistical methods seem promising, and research on effective interventions and reliable statistical adjustments will be an important task to pursue in the motivation research area in the coming years.

References

Abdelfattah, F. (2010) 'The relationship between motivation and achievement in low-stakes examinations', *Social Behavior and Personality*, Vol. 38, pp.159–168.

American Educational Research Association (AERA), American Psychological Association (APA) and National Council for Measurement in Education (NCME). (2014) *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington, DC.

Attali, Y. (2016) 'Effort in low-stakes assessment: what does it take to perform as well as in a high-stakes setting?' *Educational and Psychological Measurement*. doi:10.1177/0013164416634789 (Published online first).

Barry, C.L., Horst, S.J., Finney, S.J., Brown, A.R. and Kopp, J.P. (2010) 'Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing', *International Journal of Testing*, Vol. 10, pp.342–363.

Baumert, J. and 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*, Vol. 16, pp.441–462.

Butler, J. and Adams, R. (2007) 'The impact of student effort on the outcomes of international studies', *Journal of Applied Measurement*, Vol. 8, No. 3, pp.279–304.

Chan, D., Schmitt, N., DeShon, R.P., Clause, C.S. and Delbridge, K. (1997) 'Reactions to cognitive ability tests: the relationship between race, test performance, face validity perceptions, and test-taking motivation', *Journal of Applied Psychology*, Vol. 82, pp.300–310.

Cohen, J. (1969) *Statistical Power Analysis for the Behavioral Sciences*, Academic Press, New York.

Cole, J.S., Bergin, D.A. and Whittaker, T.A. (2008) 'Predicting student achievement for low stakes tests with effort and task value', *Contemporary Educational Psychology*, Vol. 33, pp.609–624.

DeMars, C.E. (2000) 'Test stakes and item format interactions', *Applied Measurement in Education*, Vol. 13, No. 1, pp.55–77. doi:10.1207/s15324818ame1301_3

DeMars, C.E. (2007) 'Changes in rapid-guessing behavior over a series of assessments', *Educational Assessment*, Vol. 12, pp.23–45.

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

Dotterer, A.M., McHale, S.M. and Crouter, A.C. (2009) 'The development and correlates of academic interests from childhood through adolescence', *Journal of Educational Psychology*, Vol. 101, No. 2, pp.509–519. doi:10.1037/a0013987

Duckworth, A.L., Quinn, P.D., Lynam, D.R., Loeber, R. and Stouthamer-Loeber, M. (2011) 'Role of test motivation in intelligence testing', *Proceedings of the National Academy of Sciences*, Vol. 108, No. 19, pp.7716–7720. doi:10.1073/pnas.1018601108

Eccles J., Adler T.F., Futterman R., Goff S.B., Kaczala C.M., Meece J.L. and Midgley C. (1983) 'Expectancies, values and academic behaviors', in Spence (Ed.): *Achievement and Achievement Motives*, Freeman, San Francisco, CA, pp.75–146.

Eklöf, H. (2006) 'Development and validation of scores from an instrument measuring student test-taking motivation', *Educational and Psychological Measurement*, Vol. 66, pp.643–656.

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

Eklöf, H., Pavešič, J.B. and Grønmo, L.S. (2014) 'A cross-national comparison of reported effort and mathematics performance in TIMSS Advanced 2008', *Applied Measurement in Education*, Vol. 27, pp.31–45.

Finn, B. (2015) *Measuring motivation in low-stakes assessments*, Research Report No. RR-15-19, Educational Testing Service, Princeton, NJ. doi:10.1002/ets2.12067

Hawthorne, K.A., Bol, L., Pribesh, S. and Suh, Y. (2015) 'Effects of motivational prompts on motivation, effort, and performance on a low-stakes standardized test', *Research & Practice in Assessment*, Vol. 10, pp.30–38.

Hopfenbeck, T.N. and Kjærnsli, M. (2016) 'Students' test motivation in PISA: the case of Norway', *The Curriculum Journal*, Vol. 27, No. 3, pp.406–422. doi:10.1080/09585176.2016.1156004

Immekus, J.C. and McGee, D. (2016) 'The measurement invariance of the student opinion scale across English and non-English language learner students within the context of low- and high-stakes assessments', *Frontiers in Psychology*, Vol. 7, p.1352. doi:10.3389/fpsyg.2016.01352

Kane, M. (2011) 'The errors of our ways', *Journal of Educational Measurement*, Vol. 48, pp.12–30.

Knekta, E. (2017) 'Are all pupils equally motivated to do their best on all tests? Differences in reported test-taking motivation within and between tests with different stakes', *Scandinavian Journal of Educational Research*, Vol. 61, pp.95–111. doi:10.1080/00313831.2015.1119723

Knekta, E. and Eklöf, H. (2015) 'Modeling the test-taking motivation construct through investigation of psychometric properties on an expectancy-value based questionnaire', *Journal of Psychoeducational Assessment*, Vol. 33, pp.662–673. doi:10.1177/0734282914551956.

Kornhauser, Z.G.C., Minahan, J., Siedlecki, K.L. and Steedle, J.T. (2014) *A Strategy for Increasing Student Motivation on Low-stakes Assessments*, Council for Aid to Education, New York.

Lau, A.R., Jones A.T., Anderson R.D. and Markle R.E. (2009) 'Proctors matter: strategies for increasing examinee effort on general education program assessments', *The Journal of General Education*, Vol. 58, No. 3, pp.196–217. doi:10.1353/jge.0.0045

Liu, O.L., Bridgeman, B. and Adler, R.M. (2012) 'Measuring learning outcomes in higher education: motivation matters', *Educational Researcher*, Vol. 41, No. 9, pp.352–362. doi:10.3102/0013189X12459679

Liu, O.L., Rios, J.A. and Borden, V. (2015) 'The effects of motivational instruction on college students' performance on low-stakes assessment', *Educational Assessment*, Vol. 20, pp.79–94.

Muthén, L.K. and Muthén, B.O. (1998–2012) *Mplus User's Guide*, 7th ed., Muthén & Muthén, Los Angeles, CA.

OECD (2013) *PISA 2012 Results: What Students Know and Can Do: Student Performance in Mathematics, Reading and Science*, Vol. I, OECD, Paris.

OECD (2014) *PISA 2012 Technical Report*, OECD, Paris.

O'Neil, H.F., Abedi, J., Miyoshi, J. and Mastergeorge, A. (2005) 'Monetary incentives for low-stakes tests', *Educational Assessment*, Vol. 10, No. 3, pp.185–208. doi:10.1207/s15326977ea1003_3

Penk, C., Pöhlmann, C. and 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*, Vol. 5. doi:10.1186/s40536-014-0005-4

Penk, C. and Schipolowski, S. (2015) 'Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation', *Learning and Individual Differences*, Vol. 42, pp.27–35.

Rutkowski, L. and Rutkowski, D. (2016) 'A call for a more measured approach to reporting and interpreting PISA results', *Educational Researcher*, Vol. 45, pp.252–257. doi:10.3102/0013189X16649961

Rutkowski, D. and Wild, J. (2015) 'Stakes matter: student motivation and the validity of student assessments for teacher evaluation', *Educational Assessment*, Vol. 20, No. 3, pp.165–179. doi:10.1080/10627197.2015.1059273

Schermelleh-Engel, K., Moosbrugger, H. and Müller, H. (2003) 'Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures', *Methods of Psychological Research*, Vol. 8, No. 2, pp.23–74.

Sessoms, J. and Finney, S.J. (2015) 'Measuring and modeling change in examinee effort on low-stakes tests across testing occasions', *International Journal of Testing*, Vol. 15, pp.356–388. doi:10.1080/15305058.2015.1034866

Smith, L.F. and Smith, J.K. (2002) 'Relation of test-specific motivation and anxiety to test performance', *Psychological Reports*, Vol. 91, pp.1011–1021.

Sundre, D.L. and Finney, S.J. (2002) *Enhancing the Validity and Value of Learning Assessment: Furthering the Development of a Motivation Scale*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Sundre, D.L. and 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 Pscyhology*, Vol. 29, pp.6–26.

Sundre, D.L. and Moore, D.L. (2002) 'The student opinion scale: a measure of examinee motivation', *Assessment Update*, Vol. 14, pp.8–9.

Sundre, D.L. and Wise, S.L. (2003) *Motivation Filtering: An Exploration of the Impact of Low Examinee Motivation on the Psychometric Quality of Tests*. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL, April 2003.

Swerdzewski, P.J., Harmes, J.C. and Finney, S.J. (2011) 'Two approaches for identifying low-motivated students in a low-stakes assessment context', *Applied Measurement in Education*, Vol. 24, No. 2, pp.162–188. doi:10.1080/08957347.2011.555217

Thelk, A.D., Sundre, D.L., Horst, S.J. and Finney, S.J. (2009) 'Motivation matters: using the student opinion scale to make valid inferences about student performance', *The Journal of General Education*, Vol. 58, No. 3, pp.129–151. doi:10.1353/jge.0.0047

Wentzel K.R. and Wigfield, A. (2009) 'Introduction', in Wentzel, K.R., Wigfield, A. (Eds.): *Handbook of Motivation at School*, Routledge, New York, pp.2–8.

Wigfield, A. and Eccles, J.S. (2000) 'Expectancy-value theory of achievement motivation', *Contemporary Educational Psychology*, Vol. 25, No. 1, pp.68–81.

Wigfield, A., Tonks, S. and Lutz Klauda, S. (2009) 'Expectancy-value theory', in Wentzel, K.R., Wigfield, A. (Eds.): *Handbook of Motivation at School*, Routledge, New York, pp.55–75.

Wolf, L.F. and Smith, J.K. (1995) 'The consequence of consequence: motivation, anxiety, and test performance', *Applied Measurement in Education*, Vol. 8, No. 3, pp.227–242.

Wise, S.L. and Cotton, M.R. (2009) 'Test-taking effort and score validity. The influence of student conceptions of assessment', in McInerney, D.M., Brown, G.T.L., Liem, G.A.D. (Eds.): *Student Perspectives on Assessment: What Students Can Tell Us about Assessment for Learning* (pp.187–206). Charlotte, NC: Information Age Publishing.

Wise, S.L. and DeMars, C.E. (2005) 'Low examinee effort in low-stakes assessment: problems and potential solutions'. *Educational Assessment*, Vol. 10, No. 1, pp.1–17.

Wise, S.L. and Kong, X.J. (2005) 'Response time effort: a new measure of examinee motivation in computer-based tests', *Applied Measurement in Education*, Vol. 18, pp.163–183.

Wise, S.L, Pastor, D.A. and Kong, S.J. (2009) 'Correlates of rapid-guessing behavior in low-stakes testing: implications for test development and measurement practice', *Applied Measurement in Education*, Vol. 22, pp.185–205.

Wise, S.L. and Smith, L.F. (2016) 'The validity of assessment when students don't give good effort', in Brown, G.T.L., Harris, L.R. (Eds.): *Handbook of Human and Social Conditions in Assessment*, Routledge, London, pp.204–220.

Zilberberg, A., Brown, A.R., Harmes, J.C. and Anderson, R.D. (2009) 'How can we increase student motivation during low-stakes testing? Understanding the student perspective', in McInerney, D.M., Brown, G.T.L., Liem, G.A.D. (Eds.): *Student Perspectives on Assessment: What Students Can Tell Us about Assessment for Learning*, Information Age Publishing, Charlotte, NC, pp.255–278.