TRAMES, 2019, **23(73/68)**, 3, 353–376

PREDICTING PERFORMANCE IN A LOW-STAKES TEST USING SELF-REPORTED AND TIME-BASED MEASURES OF EFFORT

Gerli Silm, Olev Must, and Karin Täht

University of Tartu

Abstract. For the validity of test results in low-stakes testing, it is important to take into account the motivation of the test-takers. Previous studies using various test-taking motivation measures have not provided coherent results. The aim of the current study was to specify the predictive power of two particular motivation indicators: self-reported effort (SRE) and response time effort (RTE). A previous high-stakes test result and gender were also added to the model to predict cognitive test performance. The sample group consisted of 280 Estonian higher education students (mean age 21.5 years (SD = 2.1), 25% male). The model was able to explain 75.6% of the variance in the test results. The predictive power of RTE was larger, but SRE supplemented the overall predictive power of the model. Using average time spent on incorrect items also proved to be a good indicator of effort.

Keywords: test-taking motivation, test-taking effort, test-taking time, response time effort (RTE), self-reported effort (SRE), motivation filtering, structural equation modelling

DOI: https://doi.org/10.3176/tr.2019.3.06

1. Introduction

In education and psychology an important part of assessing a test-taker's actual abilities is determining whether the test-taker put in their best effort on the test. The effort exerted by a test-taker is directly related to their motivation for completing the assessment and can be represented by the concept of test-taking motivation (TTM). This is defined as a goal-directed activity of "giving one's best effort to the test, with the goal being to accurately represent what one knows and can do in the content area covered by the test" (Wise and DeMars 2005:2). Test-taking effort is defined as a test-taker's "engagement and expenditure of energy toward the goal of attaining the highest possible score on the test" (Wise and DeMars 2005:2).

Over the last decades it has been established that there is a relationship between TTM and test results. Test results from low-stakes tests are on average lower than those of high-stakes tests (e.g. DeMars 2000, Duckworth, Quinn, Lynam, Loeber and Stouthamer-Loeber 2011, Napoli and Raymond 2004, Sundre 1999, Wolf and Smith 1995). In their meta-analysis Wise and DeMars (2005) found an effect size of .59 when comparing the results of motivated and less motivated test-takers. This difference can be attributed to the lower motivation of test-takers when involved in low-stakes tests. This in turn can potentially influence the validity of the inferences that are based on the test results (Haladyna and Downing 2004).

Nevertheless, research is still going on as to how TTM could be measured more precisely. At present self-report instruments are the most common measures used for determining TTM (e.g. Student Opinion Scale, Sundre and Moore 2002), but there are other approaches that can also be used for gathering TTM data, such as the use of test-taking time (e.g. Response time effort (RTE), Wise and Kong 2005) and person fit statistics (Meijer and Sijtsma 2001). The aim of the current study was to specify the predictive power of two particular TTM indicators: self-reported effort (SRE) and time-based measures of effort such as RTE.

1.1. Formation of test-taking motivation

It has been shown that performance in a cognitive test can be influenced by factors other than ability. These can include the test-taker's attitude, fatigue, anxiety, motivation, effort, as well as the test format, and the length of the test itself (e.g. Ackerman and Kanfer 2009, DeMars 2000, Duckworth et al. 2011, Silm, Must, and Täht 2013, Wolf and Smith 1995). Test-taking patterns and behavior (e.g. Must and Must 2013, Stenlund, Eklöf, and Lyrén 2017) have also been shown to have a profound impact on the actual attempts to measure individual differences in cognitive ability. Conversely, it has also been shown that the effect of motivation on academic achievement is relatively negligible when controlling for IQ (Gagné and St Père 2002, Reeve and Lam 2007). Therefore, it is important to take indicators of actual abilities (previous performance) into consideration when evaluating TTM.

The determination of test-taking motivation is usually described within the framework of the expectancy-value theory. The expectancy-value theory helps to explain why people choose to engage in, or opt out of, different activities. Eccles and Wigfield (2002, Wigfield and Eccles 2000) proposed that achievement related choices and performance are linked to an expectation of success and a subjective task value. Neither of these domains stand out on their own but are affected by previous experiences and self-schemata (see Wigfield and Eccles 2000).

1.2. Low-stakes and high-stakes testing

For the most part lack of motivation appears to be a potential problem in low-stakes testing conditions. This is in alignment with the expectancy-value theory. Quite often a low-stakes test will have no meaningful consequence for the test-taker, although it may have consequences on a super-ordinate level (e.g. international tests like the PISA (Programme for International Student Assessment)). Moreover, there are some tests that may be perceived as being of minimal importance for the test-takers, such as, for example, voluntary-based studies, pilot studies, and etc. It is precisely these kinds of tests where lack of motivation can be seen as a problem.

A test is considered high-stakes when it becomes an important factor in determining a student's (or test-taker's) future, e.g. graduation, compensation, and penalties. These kinds of tests have personal consequences for the test-taker, so it is more likely that the test-takers will have high motivation to complete this kind of a test. The difference between a high-stakes test and low-stakes test is often determined by context. Low-stakes tests are not distinguished from high-stakes tests based on their form, but rather by their function, or more precisely, how the results are used.

1.3. Self-report measures of TTM

Researchers have attempted to assess test-taking motivation using several methods. One such method is to query the test-takers directly using a self-report instrument. There are several self-report instruments that have been created for this purpose. These include: the Questionnaire of Current Motivation (Vollmeyer and Rheinberg 2006), the Motivation Questionnaire (Knekta and Eklöf 2015), the Test-Taking Motivation Questionnaire (Eklöf 2006) and the PISA effort thermometer (Kunter et al. 2002), as well as several others. The most widely used self-report instrument at present is the Student Opinion Scale (SOS; Sundre and Moore 2002).

SOS is a ten-item questionnaire comprised of two subscales. These are: the importance of the test, and the effort exerted. A two-factor structure has been confirmed by several different samples (Sundre and Finney 2002, Sundre and Thelk 2007). SOS is generally used as a post-test assessment in order to determine the test-takers' motivation in a low-stakes testing situation (Sundre and Thelk 2007). Some studies have attempted to measure TTM before the test, or even in between parts of the test (e.g. Bensley et al. 2016, Penk and Richter 2017, Penk and Schipolowski 2015, Weirich, Hecht, Penk, Roppelt, and Böhme 2017).

1.4. Time-based measures of TTM

Information about TTM, or more precisely test-taking effort, can be gathered by monitoring test-taking time (e.g. Silm, Must, and Täht 2013). This is a relatively easy task when using computer-based testing. If one person finishes a test in only 20 minutes, whereas others need an hour to complete it, then it is most probable that the person who finished quickly did not exert as much effort. A person may also skip the items that seem too difficult. Therefore, it is beneficial to start by looking at the total time a test-taker used for the test. This however does not give us any information on how the test-taker actually used their time and whether it was more or less equally distributed between all the items, or whether they spent substantially more time on one item and then skipped the rest. In these two cases the total test-taking time could actually be the same. The mean time spent on correctly answered items and on incorrectly answered items are also good indicators. The time spent on the incorrectly answered items is an especially useful indicator. Here it is possible to discern whether there was very little time on the items and whether the incorrect items might have been the result of rapid responding. However, if the test-taker spent a considerable amount of time on the incorrect items too, then it is logical to conclude that even though the item turned out to be too difficult for the respondent, they nevertheless still attempted to solve it.

A more specific measure designed to monitor test-taking time is response time effort (RTE, Wise and Kong 2005). RTE is designed to detect unmotivated test-takers who tend to answer too quickly, without fully examining the item or even properly reading it (Wise and Kong 2005: 163). RTE assigns a time threshold to each of the test items depending for example on the item's surface features (length, auxiliary materials) and the frequencies of the response times (Wise and Kong 2005). Several other methods for identifying these thresholds have also been developed (see Kong, Wise and Bhola 2007, Wise and Ma 2012). Kong et al. (2007) came to the conclusion that different threshold setting methods can still yield similar results. Wise (2017) suggested that a normative threshold approach be used to set the RTE threshold. With this method the threshold can be set to the point in time where the accuracy of the responses begins to increase from a rapid guessing rate baseline. Finding this particular point requires a large sample size. Nevertheless, Wise and Ma (2012) found that usually this point in time is approximately 10% of the mean response time of a particular item. In all cases the threshold determines the minimal time considered necessary to read the item and to give a response. If the response is given in less time than the threshold, then this is considered to be the result of careless responding. In other words, response times are compared to the thresholds in order to establish whether the answer to an item was a rapid response or not. An overall index ranging from 0 to 1 is calculated based on the determination of each response. An index of zero means that all of the responses were rapid guesses, or in other words, the response times for each of the items were below the given thresholds. An index of one means the opposite – the response times to all of the items exceeded the thresholds. For a more thorough discussion of the advantages and disadvantages of the various methods for determining RTE thresholds, see Wise (2017).

1.5. The relationship between self-reported effort and response time effort

There are few studies measuring TTM that attempt to distinguish motivated test-takers from the non-motivated by using a combination of self-reported effort (SRE), which is measured via the SOS effort subscale, in combination with RTE. Wise and Kong (2005) found that the correlation between the scores of the RTE and SOS effort subscale was .25 ($p < .001$), which was lower than expected considering that both measures are designed to gauge test-taking effort, albeit in different ways. Later studies have found the correlation between RTE and self-reported effort to be about .40, depending on the threshold setting method for RTE (Kong, Wise, and Bhola 2007), and .61 (Rios, Liu and Bridgeman 2014), although the latter study had a relatively small sample size.

According to Wise and Kong (2005) this low correlation can perhaps be attributed to the possible ceiling effect of RTE (the index was 1.0 for most of the test-takers). It is also possible that self-reported effort may be subject to different biases, such as for example (Wise and Kong 2005, Wise and DeMars 2006): 1) social desirability, wherein a higher effort than what was actually expended may be reported, 2) reporting a lower effort in order to justify a result that was lower than expected, 3) a test-taker who did not know how to estimate their effort, 4) a test-taker who did not attempt to expend effort on the test itself, nor try to thoughtfully answer the self-report questionnaire. The fact that SRE is a test-based measure, while RTE is item-based measure (Rios et al. 2014) can also have an influence. Nevertheless, self-reports may give valuable information in paper-pencil tests in which case response times are unavailable.

The low to moderate correlation between the two measures would imply that these cannot be used interchangeably. Nevertheless, Swerdzewski, Harmes, and Finney (2011) found that both of these measures can actually differentiate the non-motivated test-takers from the motivated in a quite similar manner. On the other hand, Rios et al. (2014) found that RTE has a stronger relationship with test performance and is more effective in identifying low-effort examinees. Although both measures are designed to provide information about text-taking motivation, or more precisely the specific test-taking effort of individuals, the relatively low correlations between the two measures and the contentious findings with regards to motivation filtering, means that we do not yet fully understand the underlying mechanisms for these two measures and to what extent they measure the same thing.

1.6. Improving validity

As low motivation is considered to have a negative impact on validity, measures of TTM can be applied towards improving validity. For example, person-fit statistics can be used to detect aberrant response patterns. Data from the aberrant response patterns can be then be filtered out. Although using person fit statistics has proven useful in personality research (e.g. Allik, Realo, Mõttus, Borkenau, Kuppens, and Hřebíčková 2012, Allik, Hřebíčková, and Realo 2018), at present they have not been widely used as measure of TTM. The main reason for this is that person-fit statistics can highlight different aberrant behaviours beside test-taking motivation and it is hard to conclude whether the poor fit can be attributable to a lack of effort or something else entirely (Wise and Kong 2005). For this reason, we chose not to include person-fit statistics in our study.

In TTM research motivation filtering is usually conducted using self-report measures and RTE. It has been shown that the results are significantly higher when the results of examinees with extremely low motivation (based on SRE or RTE) have been filtered out of the dataset (Wise, Wise, and Bhola 2006).

1.7. Previous empirical models of TTM

Several researchers have modeled the impact of different motivational aspects on test results and found them to be consistent with expectancy-value theory. Most of these models confirm that effort is a significant predictor of test performance, but also that it mediates some perceived value components, such as importance (Cole, Bergin, and Whittaker 2008, Knekta and Eklöf 2015, Penk and Schipolowski 2015, Zilberberg, Finney, Marsh, and Anderson 2014), usefulness (Cole et al. 2008), and expectancies (Knekta and Eklöf 2015).

There have also been several studies examining changes in TTM during the test. These studies have concluded that to some extent, TTM evolves and changes during the test-taking experience (Penk and Richter 2017, Penk and Schipolowski 2015). When TTM is measured with a self-report instrument after the test, these changes are not apparent. In contrast, RTE is a specific measure concentrating on each item of the test, and takes into account the test-taking process throughout the test.

1.8. The aim of the current research

The previous discussion makes clear that the similarities and differences of self-reported effort and time-based effort (such as RTE) are not clear. It is therefore worthwhile to make an effort to understand more about the two measures.

The main aim of the current research is to then describe and model the two effort variables. For our study this was done both separately and together, in order to see what effect the measures had on test results and how they relate to one another, as well as how they relate to other variables that are important in the context of cognitive test performance and academic achievement.

As RTE is a measure of effort, it was emphasized more in our study. RTE was then compared to SRE, which was the Effort subscale of the SOS. As there are some potential problems with the RTE index, such as the ceiling effect, we also adopted an alternative time measure in the form of average time spent on incorrect item.

1.9. Proposed model

The proposed hypothetical model for predicting test performance with self-reported effort and time-based effort is presented in Figure 1. The main emphasis of the model is placed on the two effort variables that predict performance. We also controlled for gender and previous high-stakes test results in the same manner as Cole et al. (2008), and Zilberberg et al. (2014) did in their studies. Gender differences have been found to be a factor in educational achievement (Duckworth and Seligman 2006, Pekkarinen 2012, Stoet and Geary 2015, Voyer and Voyer 2014), as well as in test-taking motivation (DeMars, Bashkov and Socha 2013) and in social desirability with regard to questionnaire responses (Bernardi 2006, Dalton and Ortegren 2011). In the case of socially desirable responding, it becomes difficult to say whether, and to what extent some people are truly more motivated, and to what extent they simply report a higher or lower motivation. It is also difficult to determine whether gender differences stem from differences in motivation, or from differences in social desirability.

As it is also known that the best predictor of achievement is a previous result from a similar task (Deary, Strand, Smith and Fernandes 2007, Goldstein 1997), it is therefore important to control for this factor in addition to TTM. This is especially important considering the findings showing that when IQ is controlled, then the effect of motivation diminishes (Gagné and St Père 2002, Reeve and Lam 2007). Penk and Richter (2017) also concluded that ideally they would have controlled for "prior knowledge with an additional measure from high-stakes test" in their model of TTM change.

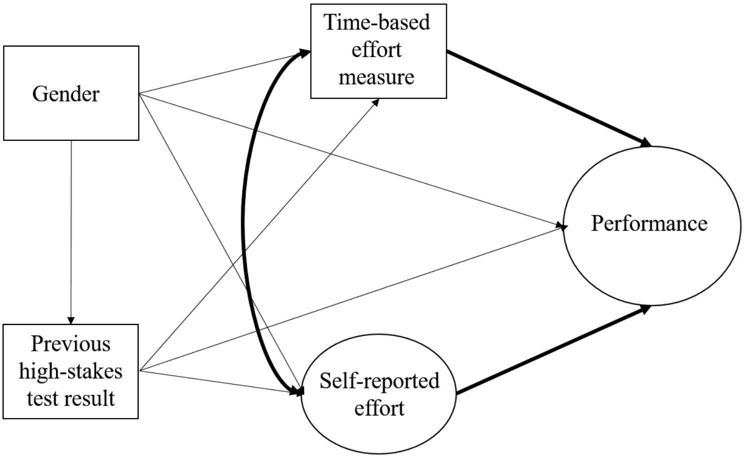

Figure 1. A hypothetical model for predicting test results.
Note: Bold lines represent relationships and effects between the main variables of the study, thin lines represent the effects from control variables.

2. Method

2.1. Sample

The study participants were first year university and college students from fourteen Estonian higher education institutions (Mägi, Adov, Täht, and Must 2013). Participation in the study was voluntary and the students were contacted via email. The study was conducted anonymously and took place in an online testing environment. Taking part in the online cognitive test had no consequences for the study participants and could therefore be considered to be a low-stakes test situation. The participants did, however, receive feedback about their results, which may have had an effect on their motivation to take the test. The feedback was provided by e-mail and included a comparison with other test-takers after the data had been gathered from all the participants.

The total sample consisted of 361 students. In our study, we used data from participants who both provided their national examination grades and filled out the SOS after taking the cognitive test. The final effective sample consisted of 280 participants. Of the final sample, 70 (25%) participants were male and 210 (75%) were female. The proportion of female test-takers was slightly higher than that of the general proportion of female students in all of the Estonian higher education establishments (60%) (Tõnisson 2011). The mean age of the sample was 21.5 years (SD = 2.1).

2.2. Instruments

2.2.1. Computer-based cognitive test

We used a shortened version of the University of Tartu's Scholastic Aptitude Test. The items were taken from several tests that had been administered between the years 2008–2012, with similar levels of difficulty ranging from .3 to .5 (Must and Allik 2002, Must and Must 2013, Silm, Must and Täht 2013). The test was made up of three subtests consisting of vocabulary, mathematics and spatial ability. In all of the subtests there were 15 items, with a total of 45 items.

2.2.2. Time-based measures of test-taking effort

We used two time-based measures of test-taking effort – RTE, and average time spent on incorrect item. To set the RTE thresholds, we used an adaption of the 10% Normative Threshold (NT10) approach (Wise and Ma 2012). This approach proposes that 10% of the average response time be used to the threshold. For our study, however, we used 10% of the average time spent on correct responses. This decision was driven by the findings that test-taking effort may decrease during the test (e.g. Weirich, Hecht, Penk, Roppelt and Böhme 2017), meaning that more rapid responses may appear in the end of the test and decreases in the average answering times may be more evident at the end of the test, rather than at the beginning. For our thresholds, we set a minimum limit of three seconds and maximum limit of ten seconds. Following the precedent of Wise and Kong (2005) we also considered the surface features approach meaning that item length and the presence of additional materials was taken into consideration when setting the RTE thresholds. In our case the average time spent on items with additional materials was similar to the average time spent on items with no additional materials. This is one reason why we opted for the NT10 approach. Also, we did not see a clear two-spike distribution in answering times. Evidence of this would have enabled the use of the visual inspection approach. On the whole, the NT10 and surface features approach seemed to produce similar results, and the resulting correlation between the RTE indexes using either approach was .995.

It should be noted that with each threshold setting method there can be classification errors (Wise 2017). These errors take the form of false positives (e.g. when effortful responses are classified as rapid guesses) or false negatives (when rapid guesses are classified as effortful responses), and the origin of these errors is not always clear. To check for these errors in our data we calculated the percentage of correct answers that were achieved using less time than the predetermined threshold. We found that in most cases (25 out of 45) there were no correct responses that were given under the threshold. In one case there was only one answer that was given with less time, and it was correct. However, when looking at other responses given by this individual, it was noted that they had received a total score of only 8 out of 45, suggesting that the correct response may have been given by chance. In other cases the percentage of correct answers that was under the threshold was either near or under the chance level (from .04 to .25 with items that had 5 response options). From this we were able to conclude that the threshold method used for our purposes was more prone to false negatives. As very short response times were indicative of rapid guessing, those test-takers who were able to quickly process the items and give the correct response were not classified as rapid guessers.

As noted previously, concern over a possible ceiling effect compelled us to look for an alternative time-based measure for test-taking effort. This led to the use of average time spent on incorrect item. For this measure we summed the response times for all incorrect responses and divided this with the number of items that were answered incorrectly or left unanswered.

Self-reported effort

SOS (Sundre and Moore 2002) is a self-reported test-taking motivation scale that consists of two subscales: Importance and Effort. The scale consists of ten questions (e.g. “Doing well in this test was important for me”, “I gave my best effort on this test”). Both subscales consisted of five questions. Test-takers filled out the questionnaire immediately after the cognitive test. The answers to the questions were based on a 5-point scale (1 – Strongly disagree, 5 – Strongly agree). For the purpose of this study the questionnaire was translated into Estonian. In our analysis we used only the effort subscale.

2.2.3. National examination results

At the end of their secondary school studies, Estonian students must typically take at least three compulsory national exams. These include an exam assessing Estonian language skills, as well as two other exams of the student’s choice from eleven that were offered in that particular year. The participants of the present study reported their achievement results from the various national exams. The mean score of these results was calculated by the authors and was then used in the analysis to serve as a benchmark for a previous high-stakes test result. The maximum possible score on a national exam is 100 points. The mean result serves as a simplified indicator, which was then used to approximately reflect the skills and abilities of the test-takers. Several researchers have adopted this approach of looking at different performance scores as one latent variable (e.g. Deary et al. 2007, Täht and Must 2010, 2013). Also it has been demonstrated that there is a high positive relationship between a previous high stakes test result and future academic achievement in different subjects (Deary et al. 2007). In their study of over 70,000 English schoolchildren, Deary et al (2007) additionally showed that IQ measured at the age of 11 has a strong correlation (r = .81) with national examination results five years later at the age of 16, thereby illustrating that national examination results are good indicators of cognitive abilities.

2.3. Research design

Our study, including the cognitive test, was Internet-based, meaning that it could be taken anywhere and at any time. Test-takers were given 60 minutes to complete the test. Immediately after, the test-taker was asked to self-report on their motivation via SOS. The test-takers were not informed that the time it took to complete the test was also being measured.

The test was conducted in a secure online environment using the university server (kaemus.psych.ut.ee). The respondents' data was anonymized for the data analysis process. In the online environment the respondents were asked to give their informed consent on the first page of the study.

Data analysis was performed using the statistical software packages SPSS Statistics 25 for descriptive and correlation analysis, and Mplus 7.2 for the structural equation modelling.

3. Results

The test-takers were given 60 minutes. The authors considered this time limit to be a quite generous, and one that would not place considerable time pressure on the test-takers. From Figure 2 it can be seen that most of the respondents did not use all the available time. The average time used for taking the test was 37.3 minutes (SD = 12.4; min = 3.1; max = 59.5). The average time spent on a correctly answered item was 52.5 seconds and the average time spent on an incorrect item was 48.8 seconds, t (279) = 3.2, p = .002. Less time spent on incorrect items suggested the presence of rapid responses.

We also looked at the individual time-use patterns of the participants who had taken the most time for the test. Almost all of them had an RTE index of over .90. There were few test-takers with a lower RTE index. One test-taker did not respond

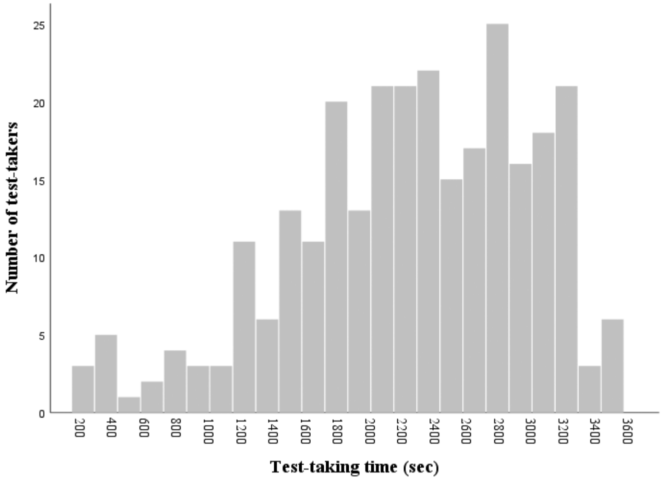

Figure 2. The distribution of total test-taking times in seconds.

to the last 10 items but spent more than 37 minutes on one item. In this case the low RTE should not be attributed to time pressure, but rather that the test-taking was interrupted by other activities. In another case the answering times were more evenly distributed – the most time spent on an item was over 7 minutes, and it was evident that for the last 5 items rapid responses were given. In this case time pressure could have been an issue. Nevertheless, there were no other tests showing a great deal of time spent on the test together with a low RTE index. In the two cases where this did occur, it is not clear whether time pressure was the agent leading to rapid responding or whether it was due to other activities that disturbed the test-taking.

The mean results and the standard deviations of the test scores and motivational indicators can be found in Table 1. In Table 2 the correlations between the two are presented.

Table 1. Descriptive statistics of used variables (N = 280)

	Vocab. (max=15)	Math. ability (max=15)	Spatial ability (max=15)	Total score (max=45)	RTE index (max=1)	Self-reported effort (max=25)	Prev. HS results (max=100)	Average time spent on incorrect item (sec)
M	8.82	9.33	8.50	26.65	.96	14.98	71.83	48.8
SD	2.87	4.18	3.69	9.17	.14	5.23	12.21	18.6
MIN	2	0	0	2	.16	5.00	30.00	6
MAX	15	15	15	43	1.00	25.00	95.00	110

Table 2. Correlations between used variables (N = 280)

	1.	2.	3.	4.	5.	6.	7.	8.
1. RTE index	1.00							
2. Vocabulary	.341**	1.00						
3. Mathematical ability	.462**	.533**	1.00					
4. Spatial ability	.533**	.503**	.688**	1.00				
5. Total score	.532**	.759**	.900**	.874**	1.00			
6. Motivation effort subscale	.167**	.191**	.348**	.392**	.377**	1.00		
7. National examination results	.227**	.547**	.517**	.449**	.587**	.118*	1.00	
8. Average time spent on incorrect item	.394**	.402**	.374**	.317**	.424**	.227**	.205**	1.00

Note **. Correlation is significant at the .001 level (2-tailed).

*. Correlation is significant at the .05 level (2-tailed).

The reliability estimates (Cronbach's alphas) for the effort measures are as follows: RTE index (whole test, N = 45) α = .974; RTE index (vocabulary, N = 15) α = .872; RTE index (mathematics, N = 15) α = .970; RTE index (spatial abilities, N = 15) α = .977; self-reported effort (N = 5) α = .897. The reliability estimates were good for both measures. Only the Cronbach's alpha for RTE within the vocabulary subtest was somewhat lower, but this may be due to the nature of the vocabulary subtest being different from the mathematics and spatial abilities subtests. It may also be that TTM was highest in the vocabulary subtest and lowest in the spatial abilities due to the fixed item order, which has been shown to have an effect on TTM (e.g. Weirich et al. 2017).

Table 2 shows that the correlation between RTE and self-reported effort was rather low (r = .17), despite both having a moderate correlation with the total test result. From Figure 3 it is evident that the relationships between SRE and the total test result, and RTE and the total result, are rather different. In both cases there is a linear relationship, but with SRE the relationship is quite dispersed, whereas RTE showed a clear ceiling effect. The relationship between the two variables is illustrated by the scatterplot in Figure 4. Here too it is possible to see the ceiling effect of RTE. Also it is apparent that some individuals self-reported a high effort, despite there being a large number of rapid responses and their receiving a rather low score on the test. When looking at the individual response patterns in this group, it is noticeable that most of the test-takers began with solution behavior, but at a certain point made the switch to rapid responding or just abandoned the test. Some test-takers used both solution behavior and rapid responding intermittently. For example, some test takers used rapid responding for the second half of mathematics subtest, but then returned to solution behavior in the beginning of spatial abilities subtest. There were however more test-takers who reported low motivation, but whose RTE score and test score were both high.

When looking at the overlap between the test-takers considered to be motivated according to RTE criteria, and those considered motivated according to self-reported effort (based on the thresholds proposed by Wise and Kong 2005 and Rios et al.

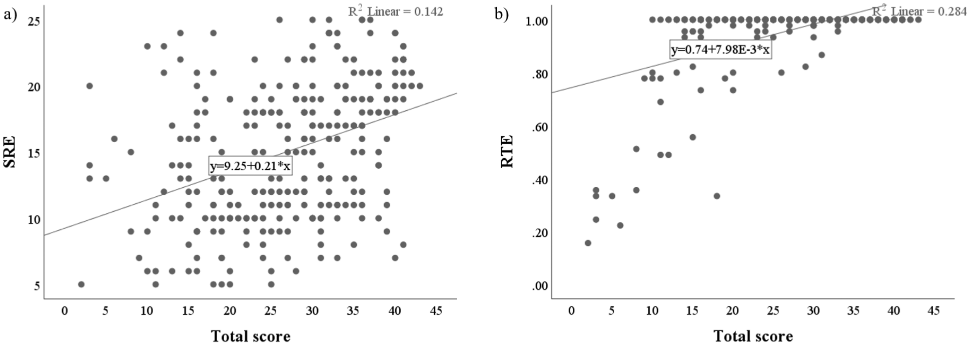

Figure 3. The relationship between SRE and total test result (a), and RTE and total test result (b).

2014; RTE ≥ 0.9, SRE ≥ 13) (Figure 4), it becomes apparent that for the most part the two overlap (57.1%). Although there are more test-takers who can be considered motivated according to the standards of RTE as compared to the self-reported effort (see Figure 5). Certainly, different thresholds can be used. This would of course change the picture, but it does indicate that RTE is able to distinguish the test-takers at the lower end of the motivation scale quite well, whereas with SRE it is not as clear.

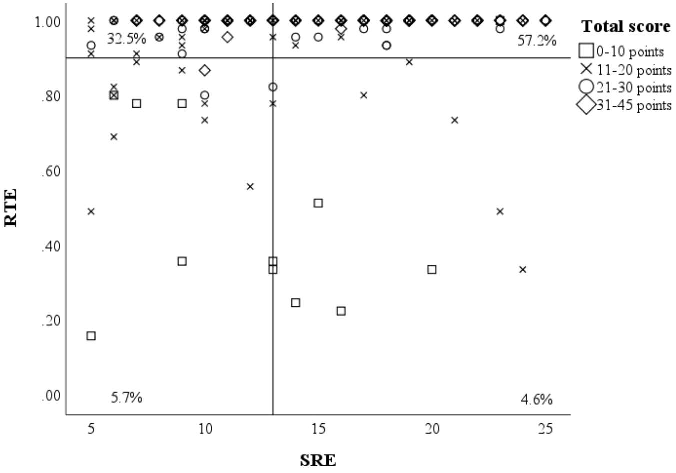

Figure 4. The relationship between SRE and RTE, and the proportions of test-takers considered motivated with either method.

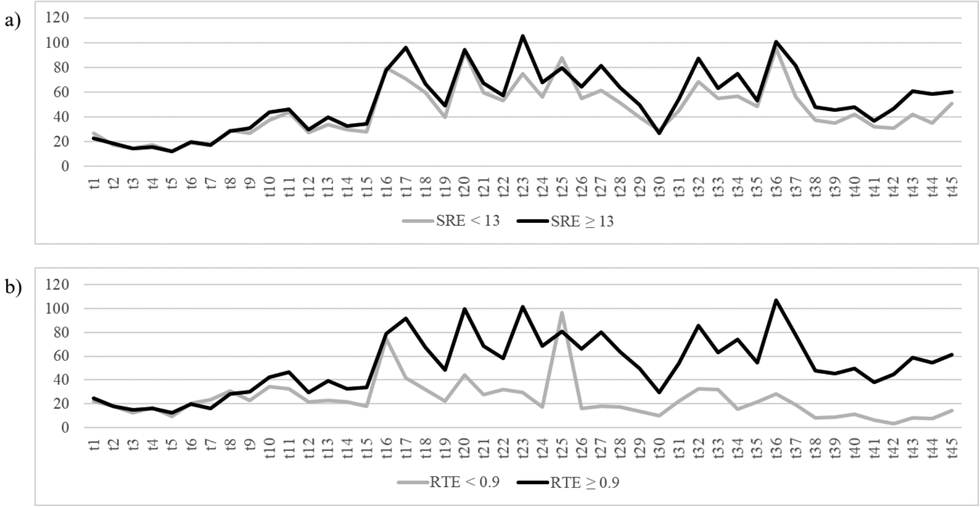

Figure 5. Average time spent on items for motivated and unmotivated test-takers based on SRE (a) and RTE (b).

It is also evident that test-takers who were considered to be motivated or unmotivated with either method, showed different patterns of time use when taking the test (see Figure 5). For most of the items, the motivated test-takers (SRE ≥ 13; RTE ≥ 0.9) spent on average more time than the unmotivated test-takers, although this is less visible in the vocabulary subtest than it is in the mathematics and spatial ability subtests.

When average time spent on incorrect item is shown, the scatterplot looks different as compared to the scatterplot with RTE (see Figure 6). There is no ceiling effect and also the correlation with self-reported effort (r = .23) is somewhat higher compared to the correlation between RTE and self-reported effort (r = .17).

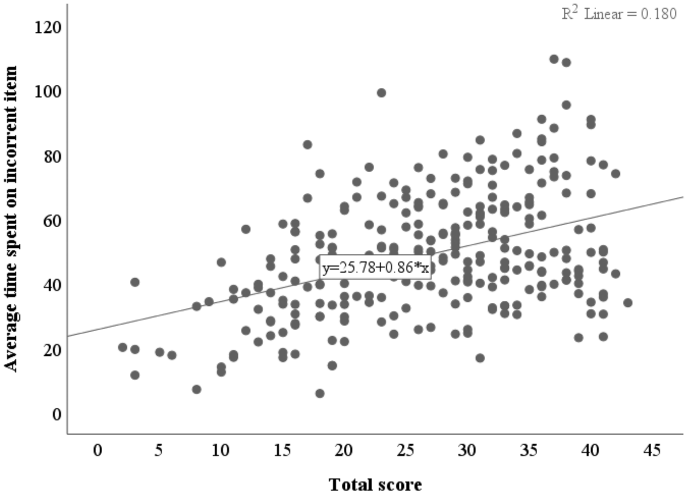

Figure 6. The relationship between SRE and average time for incorrect item (seconds).

To bring together the above findings and also incorporate other important predictors of performance, we used structural equation modeling to predict the cognitive test results. More precisely, based on our theoretical assumptions we sought to predict test performance using the two kinds of motivational indicators – time-based measure of effort (RTE or average time spent on incorrect item) and the self-reported test-taking effort. We also controlled for gender and previous high-stakes results, as these factors have been found to be related to test-taking motivation and effort.

In the model, test performance is a latent variable that impacts all three subtests of the cognitive test (this is based on the theory of *g* stating that there is a large common part in different cognitive tasks (Jensen 1998)). The self-reported effort was also modeled as a latent variable. The SEM results of the models were calculated using an MLR estimator. We chose an MLR estimator mostly because of the substantial asymmetry in RTE indicators (most values were near one, while the value of kurtosis is 16.0).

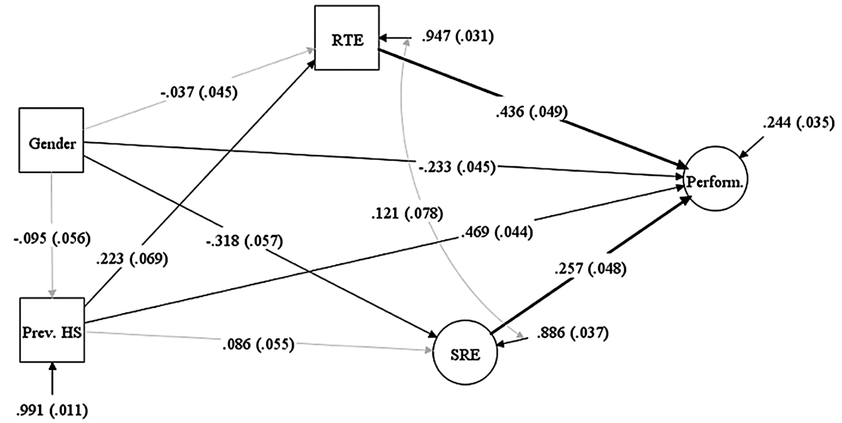

Figure 7. Model 1 for predicting test performance in low-stakes tests with self-reported effort and RTE. *Note: Perform. – test performance, RTE – response time effort, SR effort – self-reported effort, Prev. HS – previous high-stakes test result, Gender – 0-male, 1-female.*

In Model 1 (Figure 7) test performance is a latent variable that influences cognitive ability subtest scores – mathematic ability, spatial ability and verbal ability [the regression coefficients were between .632 (verbal) and .833 (spatial), while the residuals were between .600 (verbal) and .307 (spatial)]. The predictors included gender, the mean national examination result that was used as a previous high-stakes test result, and the measures of response time effort index (RTE) and self-reported effort. Self-reported effort was a latent variable that influenced five of the questions that measure effort (the regression coefficients for the effort items were between .629-.857, with residuals between .604-.265). This model predicted 75.6% of the variance of the test results.

We also created an alternative model (Model 2) using average time spent on an incorrect item, instead of RTE (see Figure 8). Here the regression coefficients for the latent performance variable were between .664 (verbal), and .838 (math), while the residuals were .558 (verbal) and .298 (math). The regression coefficients for the self-reported effort items were between .857-.627, and the residuals were .289-.607. This model predicted 65.2% of the variance in test results. Also, we found that the regression coefficient from RTE to performance (.461) were almost twice that of the regression coefficient from average time spent on incorrect item to performance (.278). In the second model the effects from the self-reported effort to performance and average time per incorrect item to performance are relatively similar with coefficients of .218 and .278 respectively.

The models ended up having quite a high predictive power. This can be explained by the addition of two motivational indicators. These motivational indicators supplemented the previous high-stakes test result and gender. The model incorporating only the previous high-stakes test result and gender explained 49.9% of the variance of the test results (the other regression paths were constrained to zero). By adding both RTE and SRE to the model it was able to improve the prediction of test result

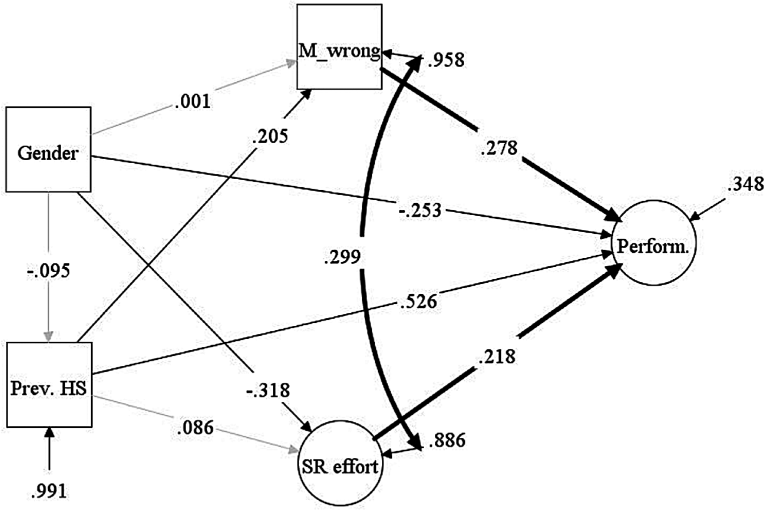

Figure 8. Model 2 for predicting test performance in low-stakes test with self-reported effort and average time spent on incorrect item. *Note: Perform. – test performance, RTE – response time effort, SR effort – self-reported effort, Prev. HS – previous high-stakes test result, Gender – 0-male, 1-female.*

by about 25%. Adding only SRE to the model improved the prediction about 8%. Adding only RTE to the model improved the prediction by about 20%. Therefore, it is evident that the different motivation indicators complement each other and have an independent effect on performance.

The following fit indices were used to evaluate the model: X^2, the root mean square error of approximation (RMSEA; Browne, Cudeck, Bollen and Long 1993), the standardized root mean square residual (SRMR; Jöreskog and Sörbom 1989), and the comparative fit index (CFI; Bentler 1990). The cut-off values developed by Brown (2006), Hu and Bentler (1999) and Dimitrov (2012) are presented in Table 3 along with the fit indices for the two models.

Table 3. Fit indices for Model 1 and Model 2 with cut-off values

	Model 1	Model 2	Cut-off values
X^2 *(df)*	102.579 (37)	115.704 (37)	
P	<.001	<.001	

RMSEA (90% confidence intervals)	.08 (.06 - .1)	.09 (.07 - .11)	≤ .05 good fit;
			≤ .08 reasonable error;
			≥ .10 poor fit (Browne et al. 1993).

SRMR	.05	.05	≤ .08 good fit (Dimitrov 2012)
CFI	.95	.94	≥ .95 good fit; .90-.95 acceptable fit (Bentler 1990, Dimitrov 2012)

It is evident from Table 3 that both models indicate an acceptable fit to the model. From the models it is also apparent that gender has an effect on performance as well as on self-reported effort, but not on time-based measures of effort. Previous high-stakes test results had the biggest effect on performance, plus on time-based measure of effort, but not on self-reported effort. Interestingly, it is also evident that with other variables accounted for, the correlation between average time spent on incorrect item and SRE is more than twice as big compared to the correlation between SRE and RTE.

4. Discussion

4.1. Predicting performance with measures of test-taking motivation

Previous studies have shown that when the two methods (SRE from the Student Opinion Scale and RTE) are used to measure TTM they tend to have a low correlation with each other. The findings in relation to motivation filtering using either of the methods have also been contradictory (Wise and Kong 2005, Kong, Wise, and Bhola 2007, Rios et al. 2014, Swerdzewski et al. 2011). The current study aimed to complement the aforementioned research by studying the effects of the two kinds of motivation indicators (SRE and RTE) on low-stakes test result. Like Rios et al. (2014) we also found that the RTE is more strongly correlated to test results, and helps to explain an additional 20% of the variance in test results when a previous result in a high-stakes test and gender have been taken into account.

More specifically, we found that when RTE was combined with self-reported effort in a path model, the measures complemented each other. However, as was expected, previous high-stakes testing result and gender described the largest amount of variance in test results. A model using only these two properties as predictors, explained 50% of the variance in test results. Adding RTE to the model increased the predictive power by 20%. Adding only SRE to the model increased the predictive power by 8%. Adding both of the motivational indicators together to the model increased the predictive power by about 26%. The alternative model that used average time spent on incorrect item instead of RTE resulted in an acceptable fit to the data. Although this model had a lower predictive power ($R^2 = .65$) than the previously described one, there was no ceiling effect as with RTE.

The fact that previous a high-stakes test result was the best predictor of test result in the model, was expected as it is well established that the best predictor of achievement is a previously measured achievement score (e.g. Goldstein 1997). But, as stated above it was evident that the motivational indicators significantly added to the prediction. Not all of the previous studies have arrived at the same result. For example Gagné and St Père (2002) and Reeve and Lam (2007) even found that the effect of motivation was negligible when controlling for IQ. In our study we also found that RTE correlated well with a previous high-stakes test result. In order to ensure that RTE is not just serving as a proxy for mental abilities Wise and Kong (2005) advised that the two should not have a significant correlation. However, nonzero relationships have been found between ability and test-taking effort (e.g. Cole et al. 2008, Penk and Schipolowski 2015, Rios, Guo, Mao and Liu, 2017, Wise, Pastor, and Kong, 2009). We would submit that in our case, the appearance of the correlation was due to the relatively difficult nature of the test.

4.2. Differences between RTE and SRE as predictors of performance

As stated above, ceiling effect for RTE was in fact evident. This may also explain the strong impact of RTE in the model. As the threshold for determining rapid responses is minimal in order to prevent the removal of correct responses that were given quickly, the RTE indexes for most of the test-takers were at their maximal. This means that the RTE index does not necessarily describe all levels of motivation, but is rather mainly representative of those who were the least motivated and tended to rely on random responses. This of course is the central aim of RTE – to detect test-takers with very low motivation. But based on RTE alone, no conclusions can be reached about the motivation of the test-takers whose response times exceeded the threshold for random responding. Here self-report, although subjective, can help to give more information, especially when the interest is to not only filter out the test-takers with very low motivation, but also to learn more about the effect of motivation on test-results. Another factor to consider with RTE is that a determination of the thresholds is often dependent on the researcher's judgement and the same threshold is then applied to all test-takers.

On the other hand, with SRE it is possible that this could have been influenced by a response bias as it was presented at the end of the test and therefore could have been affected the difficulty of the test. Also, if a test-taker used different levels of effort during the test, then it will not necessarily show up well with SRE. It was evident that some of the test-takers reported a high effort even though their tests showed a low RTE and rather low test score. When looking more closely at the response patterns of these individuals, it was apparent that they had switched to rapid responding at some point of the test.

As RTE gives more information about very low motivation, it is helpful for filtering the unmotivated test-takers from the motivated test-takers based on the thresholds proposed by Wise and Kong (2005). In our sample, which was based on these thresholds, more of the test-takers were considered motivated according to RTE criteria, but not according to the SRE. Only a few of the test-takers could be considered motivated based solely on their self-report and with the exclusion of RTE. These cases were rather atypical – either these test takers had overestimated their motivation, or as stated above, they had put considerable effort into one part of the test and then left other items unanswered or switched to rapid responding.

Interestingly, we found that gender did have an effect on the self-reported effort of the models, but not on the time-based effort indicators. This development could be attributed to the influence of social desirability when answering the self-report questionnaire. Previous studies have shown that there are gender differences in socially desirable answering (Bernardi 2006, Dalton and Ortegren 2011). Nevertheless, the test scores were also lower for the females in the sample group.

A more straightforward explanation could simply be that the actual effort applied towards the test was lower for females, just as they reported. For some reason this kind of effect was not visible for gender and RTE, nor was it visible for average time spent on an incorrect item.

4.3. Average time spent on an incorrect item as an indicator of effort

In addition to RTE, we also used another time-based measure that is designed to reflect test-taking effort: average time spent on incorrect item. We concluded that if the test-taker made an effort then they would also spend more time on the items that they found to be more difficult. A short answering time for an incorrect item would most probably indicate a rapid response. This measure could possibly be more accurate than RTE, because at some points the RTE indicator may show spurious results. For example, when a person clicked through the test, using at least 10 seconds to look at each question, he/she would be considered motivated according to the RTE index. The average time spent on incorrect item in that case would be quite low, indicating a low effort. In another case a person could have given maximal effort on two of the subtests and then left one of the subtests undone, which according to RTE index would classify them as unmotivated. Also low RTE may sometimes reflect that some items were unanswered. Unanswered items would inevitably lead to a lower score and could have inflated the relationship between RTE and test performance in the model. The same actually holds for average time spent on incorrect item when we see unanswered items as incorrect.

It has been previously concluded (see Scheerens 2016) that time spent in an educational context (time on task, learning time, engaged time) is positively related to educational achievement. It may be that some test-takers are accustomed to taking their time when they engage in more challenging tasks. As the items in the tests were quite difficult, it may also be that the ones who did not feel up to the task responded more rapidly. It has been previously found that mental taxation of the items of a test is connected to test-taking motivation (Wolf, Smith, and Birnbaum 1995, Wise and DeMars 2005). Also, the correlation between motivation and test difficulty can be connected to the expectancy of success – when a test-taker experiences success with similar tasks previously, it may contribute to their motivation to take another similar test and exert more effort. This is compatible with the expectancy-value framework.

4.4. Limitations

Using self-reported data may have been a limiting factor of the current study. Students may have been biased in their reporting of their national examination scores, or perhaps simply did not remember them. One study found that the overall validity of self-reported grades was high, although lower-achieving students were more prone to reporting their grades inaccurately (Cole and Gonyea 2010). Still, it would be useful to carry out a similar study using official records to determine students' prior academic achievement. Especially because some other studies have found that adding a previous high-stakes result could significantly contribute to further studies (e.g. Penk and Richter 2017). Another limitation of the study was related to the fact that not all of the test-takers were motivated enough to finish the test and answer the post-test questionnaire regarding their test-taking motivation. This means that there was a group of unmotivated test-takers who were left out of our sample.

The method for setting the RTE thresholds can also be considered a further limitation, as different threshold setting methods are subject to different classification errors (Wise 2017).

4.5. Further investigation

Studies, including ours, clearly demonstrate the impact of motivation, specifically test-taking effort, on low-stakes test results. Further studies should place more emphasis on clarifying the mitigating factors behind the self-reported effort and determining whether it can be attributed to social desirability, or to response bias, as we and others have suggested, or even to something else entirely.

In our study we found that the average time per incorrect item can also be a predictor of performance, but further studies are needed in order to learn if, and how, this can be used in practice, and whether a threshold can be determined based on this measure to differentiate between motivated and unmotivated test-takers. An investigation into the effect of individual profiles of test-taking patterns holds promise, as does taking into account a test taker's individualized speed of test-taking.

5. Conclusion

Our study was designed to learn more about two kinds of test-taking effort measures (time-based measures (including RTE) and self-reported effort) and their effect on test results. We were able to conclude that the two measures complement each other, and both have a unique effect on test performance even when previous high-stakes test results and gender are taken into account. In practice, it seems that it would make sense to use time-based measures of effort in addition to self-report measures whenever possible. Time-based measures have a higher predictive power in comparison to SRE. Nevertheless, SRE has a potential to differentiate between different higher levels of effort, as RTE differentiates between the ones with very low motivation.

Acknowledgments

The research has been supported by the European Regional Development Fund.

We would like to thank the anonymous reviewers for their valuable insight and suggestions that have helped to improve the manuscript.

Addresses:

Gerli Silm,
Institute of Education,
University of Tartu
Salme 1a
50103 Tartu, Estonia
E-mail: gerli.silm@ut.ee

Olev Must
Institute of Education
University of Tartu
Tartu, Estonia
E-mail: olev.must@ut.ee

Karin Täht
Institute of Psychology
University of Tartu
Tartu, Estonia
E-mail: karin.taht@ut.ee

References

Ackerman, Phillip L. and Ruth Kanfer (2009) "Test length and cognitive fatigue: an empirical examination of effects on performance and test-taker reactions". *Journal of Experimental Psychology: Applied* 15, 2, 163.

Allik, Jüri, Anu Realo, René Mõttus, Peter Borkenau, Peter Kuppens, and Martina Hřebíčková (2012) "Person-fit to the five factor model of personality". *Swiss Journal of Psychology* 71, 1, 35–45.

Allik, Jüri, Martina Hřebíčková, and Anu Realo (2018) "Unusual configurations of personality traits indicate multiple patterns of their coalescence". *Frontiers in psychology* 9, 187.

Bensley, D. Alan, Crystal Rainey, Michael P. Murtagh, Jennifer A. Flinn, Christopher Maschiocchi, Paul C. Bernhardt, and Stephanie Kuehne (2016) "Closing the assessment loop on critical thinking: the challenges of multidimensional testing and low test-taking motivation". *Thinking Skills and Creativity* 21, 158–168.

Bentler, Peter M. (1990) "Comparative fit indexes in structural models". *Psychological bulletin* 107, 2, 238.

Bernardi, Richard A. (2006) "Associations between Hofstede's cultural constructs and social desirability response bias". *Journal of Business Ethics* 65, 1, 43–53.

Brown, Timothy A. (2014) *Confirmatory factor analysis for applied research*. New York and London: The Guilford Press.

Browne, Michael W. and Robert Cudeck (1993) "Alternative ways of assessing model fit". In Kenneth A Bollen and J Scott Long, eds. *Testing structural equation models*, 136–162. (Sage focus editions, 154.) Newbury Park: Sage Publications.

Cole, James S. and Robert M. Gonyea (2010) "Accuracy of self-reported SAT and ACT test scores: implications for research". *Research in Higher Education* 51, 4, 305-319.

Cole, James S., David A. Bergin, and Tiffany A. Whittaker (2008) "Predicting student achievement for low stakes tests with effort and task value". *Contemporary Educational Psychology* 33, 4, 609–624.

Dalton, Derek and Marc Ortegren (2011) "Gender differences in ethics research: the importance of controlling for the social desirability response bias". *Journal of Business Ethics* 103, 1, 73–93.

Deary, Ian J., Steve Strand, Pauline Smith, and Cres Fernandes (2007) "Intelligence and educational achievement". *Intelligence* 35, 1, 13–21.

DeMars, Christine E. (2000) "Test stakes and item format interactions". *Applied Measurement in Education* 13, 1, 55–77.

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

Dimitrov, Dimiter M. (2014) *Statistical methods for validation of assessment scale data in counseling and related fields*. John Wiley & Sons.

Duckworth, Angela Lee and Martin E. P. Seligman (2006) "Self-discipline gives girls the edge: gender in self-discipline, grades, and achievement test scores". *Journal of educational psychology* 98, 1, 198.

Duckworth, Angela Lee, Patrick D. Quinn, Donald R. Lynam, Rolf Loeber, and Magda Stouthamer-Loeber (2011) "Role of test motivation in intelligence testing". *Proceedings of the National Academy of Sciences* 108, 19, 7716–7720.

Eccles, Jacquelynne S. and Allan Wigfield (2002) "Motivational beliefs, values, and goals". *Annual review of psychology* 53, 1, 109–132.

Eklöf, Hanna (2006) "Development and validation of scores from an instrument measuring student test-taking motivation". *Educational and Psychological Measurement* 66, 4, 643–656.

Gagné, Françoys and François St Père (2001) "When IQ is controlled, does motivation still predict achievement?". *Intelligence* 30, 1, 71–100.

Goldstein, Harvey (1997) "Methods in school effectiveness research". *School effectiveness and school improvement* 8, 4, 369–395.

Hu, Li-tze and Peter M. Bentler (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.

Jensen, Arthur R. (1998) The G factor: the science of mental ability. Westport, CT and London: Praeger.

Jöreskog, Karl G. and Dag Sörbom (1989) *LISREL 7: A guide to the program and applications.* SPSS.

Knekta, Eva (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* 61, 1, 95–111.

Knekta, Eva and Hanna Eklöf (2015) "Modeling the test-taking motivation construct through investigation of psychometric properties of an expectancy-value-based questionnaire". *Journal of Psychoeducational Assessment* 33, 7, 662–673.

Kong, Xiaojing J., Steven L. Wise, and Dennison S. Bhola (2007) "Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behaviour". *Educational and Psychological Measurement* 67, 4, 606–619.

Kong, Xiaojing J., Steven L. Wise, and Dennison S. Bhola (2007) "Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior."*Educational and Psychological Measurement* 67, 4, 606–619.

Kunter, Mareike, Gundel Schümer, Cordula Artelt, Jürgen Baumert, Eckhard Klieme, Michael Neubrand, Manfred Prenzel, et al. (2002) *PISA 2000: dokumentation der erhebungsinstrumente.* Max-Planck-Institiut für Bildungsforschung.

Mägi, Mari-Liis, Liina Adov, Karin Täht, and Olev Must (2013) "Who is willing to take low-stakes assignments?". *Trames* 17, 4, 417–432.

Meijer, Rob R. and Klaas Sijtsma (2001) "Methodology review: evaluating person fit." *Applied psychological measurement* 25, 2, 107–135.

Must, Olev and Aasa Must (2013) "Changes in test-taking patterns over time". *Intelligence* 41, 6, 780–790.

Must, Olev and Jüri Allik (2002) *Tunne oma võimeid: abivahend eneseanalüüsiks*. [Know your abilities: a tool for self-analysis.] Tartu: Tartu University Press.

Napoli, Anthony R. and Lanette A. Raymond (2004) "How reliable are our assessment data? A comparison of the reliability of data produced in graded and un-graded conditions". *Research in Higher Education* 45, 8, 921–929.

Pekkarinen, Tuomas (2012) *Gender differences in education*. Paper prepared for the Nordic Economic Policy Review Conference in Oslo, 24 October 2011. (IZA Discussion Paper Series, 6390.) Bonn: Institute for the Study of Labor (IZA).

Penk, Christiane and Dirk Richter (2017) "Change in test-taking motivation and its relationship to test performance in low-stakes assessments". *Educational Assessment, Evaluation and Accountability* 29, 1, 55–79.

Penk, Christiane and Stefan Schipolowski (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.

Reeve, Charlie L. and Holly Lam (2007) "Consideration of g as a common antecedent for cognitive ability test performance, test motivation, and perceived fairness". *Intelligence* 35, 4, 347–358.

Rios, Joseph A., Hongwen Guo, Liyang Mao, and Ou Lydia Liu (2017) "Evaluating the impact of careless responding on aggregated-scores: to filter unmotivated examinees or not?". *International Journal of Testing* 17, 1, 74–104.

Rios, Joseph A., Ou Lydia Liu, and Brent Bridgeman (2014) "Identifying low-effort examinees on student learning outcomes assessment: a comparison of two approaches." *New Directions for Institutional Research* 161, 69–82.

Scheerens, Jaap (2016) Educational effectiveness and ineffectiveness: a *critical review of the knowledge base*. Dordrecht, Heidelberg, New York, and London: Springer.

Silm, Gerli, Olev Must, and Karin Täht (2013) "Test-taking effort as a predictor of performance in low-stakes test." *Trames* 17, 4, 433–448.

Stenlund, Tova, Hanna Eklöf, and Per-Erik Lyrén (2017) "Group differences in test-taking behaviour: an example from a high-stakes testing program". *Assessment in Education: Principles, Policy & Practice* 24, 1, 4–20.

Stoet, Gijsbert and David C. Geary (2015) "Sex differences in academic achievement are not related to political, economic, or social equality". *Intelligence* 48, 137–151.

Sundre, Donna L. (1999) "Does examinee motivation moderate the relationship between test consequences and test performance?". In *Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada*.

Sundre, Donna L. and A. D. Thelk (2007) *The Student Opinion Scale (SOS). a measure of examinee motivation. Test manual*. Harrisonburg: Center for Assessment and Research Studies, James Madison University.

Sundre, Donna L. and Deborah L. Moore (2002) "The Student Opinion Scale: a measure of examinee motivation". *Assessment Update* 14, 1, 8-9.

Sundre, Donna L. and Sara J. Finney (2002) "Enhancing the validity and value of learning assessment: furthering the development of a motivation scale". In *Annual Meeting of the American Educational Research Association, New Orleans.*

Swerdzewski, Peter J., J. Christine Harmes, and Sara J. Finney (2011) "Two approaches for identifying low-motivated students in a low-stakes assessment context". *Applied Measurement in Education* 24, 2, 162–188.

Täht, Karin and Olev Must (2010) "Are the links between academic achievement and learning motivation similar in five neighbouring countries?". *Trames* 14, 3, 271.

Täht, Karin and Olev Must (2013) "Comparability of educational achievement and learning attitudes across nations". *Educational Research and Evaluation* 19, 1, 19–38.

Tõnisson, Eve (2011) *Kõrghariduse valdkonna statistiline ülevaade– 2011*. [Statistical review of higher education – 2011.] Available online at <http://dspace.ut.ee/handle/10062/40767>. Accessed on 23.07.2019.

Vollmeyer, Regina and Falko Rheinberg (2006) "Motivational effects on self-regulated learning with different tasks". *Educational Psychology Review* 18, 3, 239–253.

Voyer, Daniel and Susan D. Voyer (2014) "Gender differences in scholastic achievement: a meta-analysis". *Psychological bulletin* 140, 4, 1174.

Weirich, Sebastian, Martin Hecht, Christiane Penk, Alexander Roppelt, and Katrin Böhme (2017) "Item position effects are moderated by changes in test-taking effort". *Applied psychological measurement* 41, 2, 115–129.

Wigfield, Allan and Jacquelynne S. Eccles (2000) "Expectancy – value theory of achievement motivation". *Contemporary educational psychology* 25, 1, 68–81.

Wise, Steven L (2017) "Rapid-guessing behavior: its identification, interpretation, and implications." *Educational Measurement: Issues and Practice* 36, 4, 52–61.

Wise, Steven L. and Christine E. DeMars (2005) "Low examinee effort in low-stakes assessment: problems and potential solutions". *Educational assessment* 10, 1, 1–17.

Wise, Steven L. and Christine E. DeMars (2006) "An application of item response time: the effort-moderated IRT model". *Journal of Educational Measurement* 43, 1, 19–38.

Wise, Steven L. and Lingling Ma (2012) "Setting response time thresholds for a CAT item pool: the normative threshold method." In *Annual Meeting of the National Council on Measurement in Education, Vancouver, Canada.*

Wise, Steven L. and Xiaojing Kong (2005) "Response time effort: a new measure of examinee motivation in computer-based tests". *Applied Measurement in Education* 18, 2, 163–183.

Wise, Steven L., and Lingling Ma (2012) "Setting response time thresholds for a CAT item pool: the normative threshold method." In *Annual Meeting of the National Council on Measurement in Education, Vancouver, Canada*.

Wise, Steven L., Dena A. Pastor, and Xiaojing J. Kong (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.

Wise, Vicki L., Steven L. Wise, and Dennison S. Bhola (2006) "The generalizability of motivation filtering in improving test score validity." *Educational Assessment* 11, 1, 65–83.

Wolf, Lisa F. and Jeffrey K. Smith (1995) "The consequence of consequence: Motivation, anxiety, and test performance". *Applied Measurement in Education* 8, 3, 227–242.

Wolf, Lisa F., Jeffrey K. Smith, and Marilyn E. Birnbaum (1995) "Consequence of performance, test, motivation, and mentally taxing items". *Applied Measurement in Education* 8, 4, 341–351.

Zilberberg, Anna, Sara J. Finney, Kimberly R. Marsh, and Robin D. Anderson (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.