This is the published version of a paper published in *Assessment in education: Principles, Policy & Practice*.

Citation for the original published paper (version of record):

Group Differences in Test-Taking Behaviour: An example from a High-Stakes Testing Program.
https://doi.org/10.1080/0969594X.2016.1142935

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-117617
Group differences in test-taking behaviour: an example from a high-stakes testing program

Tova Stenlund\textsuperscript{a,b}, Hanna Eklöf\textsuperscript{a} and Per-Erik Lyrén\textsuperscript{b}

\textsuperscript{a}Department of Psychology, and Department of Applied Educational Science, Umeå University, Umeå, Sweden; \textsuperscript{b}Department of Applied Educational Science, Umeå University, Umeå, Sweden

\textbf{ABSTRACT}

This study investigated whether different groups of test-takers vary in their reported test-taking behaviour in a high-stakes test situation. A between-group design (N = 1129) was used to examine whether high and low achievers, as well as females and males, differ in their use of test-taking strategies, and in level of reported test anxiety and motivation. The results showed differences between high and low achievers on a number of test-taking strategies, where high achievers reported using successful strategies to a higher extent. There were also gender differences: females, for example, reported using random guessing to a higher extent than males. Further, low achievers, especially females, reported significantly higher levels of test anxiety than high achievers, and high achievers reported slightly higher levels of motivation when compared to low achievers. To conclude, test-taking behaviour might bring additional variance to test scores, whether or not this is irrelevant variance is discussed.

It is a desired feature of any testing program that test-takers’ performance on the test is mainly affected by the content and difficulty of the test and the competencies of the test-takers. However, how test-takers behave, feel and respond to test situations might also contribute significantly to their performance. For example, to be successful when taking a test it is important to know and use appropriate test strategies, including adequate preparation (Bicak, 2013; Bond & Harman, 1994). Further, it is important to be motivated, and to be able to suppress test anxiety and other emotions which can interfere with test performance (Naylor, 1997; Sternberg, 1998). Test-taker behaviour, such as test-taking strategies, and emotional and motivational concerns related to test-taking, may introduce construct-irrelevant variance into the test scores (Haladyna & Downing, 2004). Thus, the test may measure more than the test developers intended (i.e. not only the proficiency of the test-taker, but also how he or she manages to cope with the test situation as such). As validity refers to ‘the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests’ (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), construct-irrelevant
variance is especially important to examine in situations where the proposed uses of test scores involve high-stakes decisions for the test-takers. High-stakes tests can result in consequences that are highly valued for the test-taker, for example, to gain access to higher education or to get employment. In recent years, the interest in test-taking behaviour and how this interacts with test performance has increased; however, much remains to be done in this area (Cohen, 2006; Ellis & Ryan, 2003; Hong, Sas, & Sas, 2006). The present study focuses on group differences in test-taking strategies, including test anxiety and test-taking motivation, in a high-stakes test situation.

Test-taking strategies have been defined as rules of procedures used for understanding and solving test items (Ellis & Ryan, 2003), a set of skills which a test-taker can use to improve a test score regardless of content area (Dolly & Williams, 1986). Test-taking strategies are also said to be cognitive abilities that allow test-takers to deal with test situations in an appropriate manner and to know how to prepare, what to do during and after the test (Dodeen, 2008; Dodeen, Abdelfattah, & Alshumrani, 2014). Effective preparation includes strategies such as practice on items similar to those in the test, or to participate in test preparation workshops (Bond & Harman, 1994). Effective during-test strategies include skills related to time management, such as not spending too much time on any one item and to use the full time allowed (Ellis & Ryan, 2003). During-test strategies also include organisational strategies, such as sequence items and skip difficult items (Hong et al., 2006). Another type of during-test strategies is test-wiseness, a set of skills used when the test-taker does not know the answer. For example, to use elimination processes or look for clues in the item stem (Mahamed, Gregory, & Austin, 2006). This type of strategies is particularly applicable in multiple-choice tests. Bond (1989), who used the terms test-wiseness and test-sophistication to describe all test-taking strategies used during the test, also included skills such as being able to reduce test-anxiety and increase confidence during the test. Test-wise individuals also tend to evaluate their performance after the test and try to improve preparation methods based on the current test results (Dodeen, 2008; Hong et al., 2006). In this study, we examine group differences in terms of achievement level and gender, in test-taking strategies employed before and during a high-stakes multiple-choice test (a college admission test). Much related research was made in the second half of the twentieth century, and even though some studies have been made since then, there seems to be a lack of contemporary research in this area. However, some of the more recent studies are presented below.

Several studies have shown that high and low achievers might differ in their use of test-taking strategies (Hong et al., 2006; Kim & Goetz, 1993; McClain, 1983; Rindler, 1980). Hong et al. (2006) examined differences in use of test-taking strategies among high and low mathematics achievers. They found that low achievers tend to be more passive regarding the use of their test-taking strategies compared to high achievers. That is high achievers’ reported use of strategies, such as assessing item difficulty or sequencing items before they begin, and test-wise skills, more often than low-achieving students. However, they also found that some test-taking strategies, such as checking the correctness of answers and rereading or redoing problems, were used by high as well as low achievers. In addition, Kim and Goetz (1993), also examining high and low mathematical achievers, found that high achievers used test-wiseness skills such as skipping difficult items, elimination strategies, (i.e. analysing and dismissing the wrong answer), and anticipating answers to multiple-choice items before reading the alternatives to a higher degree than low achievers. These results were in line
with McClain (1983), who studied the exam-taking strategies of high- and low-achieving psychology students. Rindler (1980) examined the effects of skipping over more difficult items among psychology students and found that any benefit of initially skipping over an item and returning to it later was related to the students’ ability level. It seems to be more beneficial for high achievers than for low achievers. Among the low achievers, those who did not skip items performed better.

A few of the more recent studies also includes research about ethnical and gender differences in test-taking strategies (Baldiga, 2014; Ellis & Ryan, 2003). In an attempt to explain the difference in cognitive-ability test performance between African-Americans and White Americans, with African-Americans being less successful, Ellis and Ryan (2003) examined the use of test-taking strategies. They found that African-Americans reported using more ineffective test-taking strategies, such as picking an answer option at random when you guess, reading through the entire test before you start, or reading the answer option before the question. Their conclusion was that a portion of the difference in performance could be attributed to the use of ineffective strategies. The use of test-taking strategies has also been examined in attempts to explain why females generally show poorer performance in high-stakes contexts compared to men. Baldiga (2014) examined gender differences in skipping items instead of guessing. The results showed that females have a tendency to be less willing to guess on multiple choice items (when there is a penalty for wrong answers) compared to males. This result is in line with earlier research which also showed that men are more likely to guess than to skip items of which they were unsure even though no penalty was incurred for an incorrect answer (Ben-Shakhar & Sinai, 1991; Hirschfeld, Moore, & Brown, 1995).

To be successful in a high-stakes (or any) test situation, sufficient knowledge about the contents of the test and the use of effective test-taking strategies are not the only concerns. Emotional and motivational factors, such as test anxiety and test-taking motivation, may also have an impact on test performance. Test anxiety can cause a number of physiological, emotional and behavioural responses, and a load on the cognitive system, as test anxiety often includes worries and irrelevant thoughts (Carter et al., 2008). It follows that high levels of test anxiety may interfere with an optimal test performance and thus, being able to suppress, reduce or cope with anxiety is crucial for the test-takers performance. Even though there may be situations in which anxiety enhances performance (Naylor, 1997), worry during tests has been shown to correlate negatively with performance (Cassady & Johnson, 2002; Dodeen, 2009; Hancock, 2001; Naylor, 1997). Studies have also shown that there seems to be gender differences in reported anxiety, where females tend to experience higher anxiety than males (Cassady & Johnson, 2002; Naylor, 1997). Further, studies have shown that test anxiety might be related to whether a test is considered to be difficult or not (Hong et al., 2006). Predicted difficulty as a source of anxiety might be due to a lack of preparation or as a consequence of low ability. Motivational concerns are also important in test situations, as a low level of motivation to do one’s best on a given test supposedly can reduce the performance level (van Barneveld, 2007; Wise & DeMars, 2005). This can be particularly problematic in low-stakes contexts, where the test result has no consequences for the test-takers. In high-stakes contexts, the main assumption is that the level of motivation among test-takers is on average high due to the anticipated reward, such as gaining access to higher education. Still, Cheng et al. (2014) argue that motivation to succeed in high-stakes contexts is not static, but varies in relation to the test-taker and contexts based on both the intended and unintended test use. Moreover, it has been found that males and females
tend to vary in perceived motivation, with males reporting lower levels of motivation on low-stakes tests (DeMars, Bashkov, & Socha, 2013).

In sum, these studies seem to indicate that there may be differences between groups in use of test-taking strategies, as well as in test anxiety and motivation. It is, however, important to further examine these differences, especially in high-stakes tests, for several reasons. The first and most important reason is that it could bring construct-irrelevant variance into the test scores. Thus, the test scores might not only reflect the desired competencies of the test-taker but also to what extent the test-taker is able to use successful test-taking strategies, including dealing with emotional and motivational concerns in a proper manner. An often discussed issue, especially in the context of standardised high-stakes tests, is whether certain groups of test-takers are disadvantaged by the characteristics of the test. It is possible that the items in the test are not biased per se, but that differences in test-taking behaviour might be the explanation to why some groups perform poorer than others. A second reason is the educational implications group differences in test-taking behaviour might bring. For example, it might bring a need to put more effort into educating test-takers about how to use test-taking strategies, and how to deal with test anxiety and test-motivation, to achieve more unbiased scores.

**Purpose**

The purpose of the present study was to investigate whether certain groups differ in their reported test-taking behaviour during a high-stakes test situation. More specifically this study aimed at examining differences between high and low performers, as well as between females and males, in reported use of test-taking strategies, test anxiety and motivation in relation to the Swedish scholastic aptitude test (SweSAT).

The SweSAT is a paper-and-pencil, all multiple-choice test used for selection (not eligibility) to higher education in Sweden. The test is used for selection in all subject areas, and, consequently, measures what is considered to be general abilities and knowledge. Sweden has a highly centralised admissions system, and the admissions regulations stipulate that in situations where there are more applicants than available places at a certain course or program, at least one third of the candidates should be admitted on the basis of their SweSAT score and the upper secondary school GPA, respectively, and at most one-third should be admitted on the basis of what is sometimes referred to as ‘local criteria,’ that is, the institutions can set these selection criteria themselves. This means that it is not mandatory to take the SweSAT, and the test is generally viewed as a second chance for those with uncompetitive GPAs.

The test has two separately scored sections: a verbal section and a quantitative section, with 80 items each. The division of the test in two sections was introduced in 2011 after a major revision of the test, which included more items in total (from 122 to 160), three new subtests (currently eight in total) and less emphasis on vocabulary knowledge (from about one-third to one-eighth of the test items). The raw scores (number correct) on each section are transformed to corresponding normed scores, that is, verbal (V) and quantitative (Q). The total score (T) is the average of the V and Q scores, and this is the only score that is currently used for selection.

The SweSAT is administered twice each year (spring and autumn), and since the revised test was introduced in late 2011 there has been between 40,000 and 80,000 test-takers at each
administration. There is usually a rather large proportion of repeaters; of those who participated in the autumn 2012 test form 66% took the test for the first time (first-timers), 24% took it for the second time and 10% had taken the test at least two times before. There are generally large and significant score differences between first-timers and repeaters (see e.g. Törnkvist & Henriksson, 2006). First-timers at the autumn 2012 test had a mean raw total score of 76, while the corresponding score for those who took the test for the second time was 84 (a difference of 0.35 SDs compared to first-timers), and for those who took the test for the third time it was 91 (a difference of 0.65 SDs compared to first-timers). There are also other consistent group differences, for example, regarding gender. Males score higher than females on both the verbal section (Cohen's $d = 0.2$) and the quantitative section ($d = 0.5$). As these differences are not limited to SweSAT, and there are no reasons to believe that males as a group have better knowledge of the contents and skills assessed in the test, it seems highly relevant to examine gender differences in terms of reported test-taking behaviour. It should be noted that test-takers will have studied under somewhat different curricula in upper secondary school and there are differences in terms of the gender distribution at the various programs (e.g. more male test-takers from technical or natural science-oriented programs, and more female test-takers from social science or performing arts oriented programs). However, there are differences in SweSAT scores between men and women even at specific programs, such as the natural science program and the social science program.

**Method**

**Participants and grouping procedures**

The sample for this study consisted of individuals that had registered for the SweSAT in the autumn of 2012. The questionnaire was sent out by email to a random sample ($n = 5468$) of these registrees. Of the 1926 (35% of the total sample) that responded to the questionnaire, 141 did not take the SweSAT at all and 38 left off the test before completion or took the test with accommodations (such as dyslectics with extended time). Further, of the 1747 who took the SweSAT without accommodations, 618 (35.3%) were repeaters. Only data from the respondents who completed the SweSAT for the first time and without accommodations were considered to be relevant for this study ($n = 1129$). This group consisted of 63.2% females and the mean age was 22 (SD = 6.7).

In this sample, there is a statistically significant difference between females’ and males’ total raw score (number correct) on the SweSAT, $t (796.4) = 5.96$, $p < .001$, with males receiving higher scores ($M = 86.0$, SD = 24.6) than females ($M = 77.5$, SD = 22.2). There was no significant difference between females and males in age $t (1127) = .45$, $p = .652$.

**Grouping procedure of high and low achievers**

As it is known that using median or mean split to differentiate groups on any continuous variable might increase the chance of finding spurious effects (see e.g. Maxwell & Delaney, 1993), we have used percentiles to ensure that the respondents/test-takers indeed are distinguished based on their achievement level, see the criteria below. Note that percentiles and scores are for all first time SweSAT test-takers in the autumn 2012, not just for the test-takers that responded to the questionnaire.
• **Low achievers.** To be considered a low achiever, a test-taker had to be ranked at or below the 40th percentile which corresponds to a raw score of 68 or lower. This group \((n = 377, 33.4\% \text{ of the sample})\), had 71% females, and the mean age was 20.8 (SD = 5.2).

• **High achievers.** To be considered a high achiever, a test-taker had to be ranked at the 60th percentile or above, which corresponds to a raw score of 80 or higher. This group \((n = 546, 48.4\% \text{ of the sample})\) had 56% females, and the mean age was 22.9 (SD = 6.8).

There is a statistically significant difference between high and low achievers in age, \(t(921) = -5.01, p < .001\), and gender, \(x^2 (1) = 22.76, p < .001\).

The moderate group, those who ranked between the 41st and the 59th percentile (scores between 72 and 79; \(n = 206 \text{ (18.2\% of the sample)}\)), had 67% females, mean age 21.8 (SD = 6.23), is not used in the analysis regarding differences between high and low achievers. We are aware of the loss of statistical power when splitting the sample and reducing the sample size with about 20%. Yet, the large total sample size in this study makes it possible to do this and still have adequate statistical power.

**Grouping procedure II, extreme high and low achievers**

To examine whether the results of this study differ when only including the extreme high and low achievers, thus even further differentiating the respondents/test-takers based on their achievement level, all analyses were repeated only including the extremes. The following criteria were used in this second grouping procedure.

• **Extreme low achievers.** To be considered an extreme low achiever, a test-taker had to be ranked at or below in the 20th percentile for all first time test-takers in the autumn of 2012 \((n = 171, 15\%)\), i.e. scored lower or equal to 55 on the SweSAT. This group consisted of 123 females (72%) and 48 males. Mean age 20.6 (SD = 4.6).

• **Extreme high achievers.** To be considered an extreme high achiever, a test-taker had to be ranked at or below the 80th percentile for all first-time test-takers in the autumn of 2012 \((n = 316, 28\%)\), i.e. scored equal to or higher than 95 on the SweSAT. This group consisted of 165 females (52%) and 151 males. Mean age 23.6 (SD = 8.5).

**Procedure and instrument**

The data used in the present study were collected through a post-test, self-report questionnaire used in the autumn 2012 administration of the SweSAT. The questionnaire was web-based and it was open from three days after the SweSAT test day until two days before the test-takers could get their online score report, a period of four weeks. Two reminders were sent out, the first after six days and the second after another six days. In all, we obtained a 35% response rate, a less than optimal proportion but nevertheless rather typical for online surveys (see e.g. Nulty, 2008). Compared to the total sample of test-takers, there was an overrepresentation of females in the responding sample, but similar age and achievement score distributions. Given the purpose and design of the present study, it was decided to use the obtained data without further adjustment.

The purpose of this questionnaire was to measure test-takers’ perceptions of different aspects of the SweSAT (e.g. difficulty, relevance, timing issues and the current revisions) as well as test-taking strategies, emotional and motivational concerns from the test-takers’
perspective. As the focus of this study is on test-taking behaviour and other non-cognitive variables that may be important for test performance, the part of the questionnaire evaluating the revised SweSAT is not relevant for this study and therefore not reported in this text.

**Test-taking strategies**

The part of the questionnaire examining specific test-taking strategies was divided into test-preparation strategies and during-test strategies.

The test-preparation section included six yes/no items about, for example, whether the test-takers had practiced on previously administered test forms, or if they had participated in workshops or studied content areas included in the test (see Table 1 for a more specific description). The during-test strategies section consisted of 12 items which were rated on a 5-point scale (from strongly disagree to strongly agree). The items were divided into three groups of strategies; structural organisation, time management and test-wiseness. The categorisation of these strategies was simply based on item content. See Table 2 for specific information about the items in each group. The items asking for test-taking strategies were

| Table 1. Frequency and per cent use of the different preparation strategies divided by group. |
|---------------------------------|---------------------------------|
| **Test-preparation strategies**  | **Low achievers** | **High achievers** |
| Use (n) | % | Use (n) | % |
| 1. Participated in a course or workshop | 11 | 2.9 | 13 | 2.4 |
| 2. Studied the areas included in the test | 117 | 31.0 | 175 | 32.1 |
| 3. Practice old tests with time limits | 85 | 22.5 | 148 | 27.1 |
| 4. Practice old tests without time limits | 135 | 35.8 | 205 | 37.5 |
| 5. Viewing old tests to get an idea of the content | 211 | 56.0 | 269 | 49.3 |
| 6. Information about the test and its content areas | 215 | 57.0 | 381 | 69.8 |

*p-value < .05.
**p-value < .01.

| Table 2. Frequency (N) Mean (M), and standard deviation (SD) of reported use of during-test strategies among high and low achievers. |
|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| **During-test strategies**  | **Low achievers** | **High achievers** |
| **N** | **M** | **SD** | **N** | **M** | **SD** |
| Structural organisation | | | | | | |
| 1. I looked through the test before I started | 363 | 1.87 | 1.21 | 542 | 1.61 | 1.03 |
| 2. I answered the items in the order they were displayed in the test booklet | 364 | 4.15 | 1.17 | 543 | 4.00 | 1.21 |
| 3. I jumped between items and subtests | 358 | 1.70 | 1.05 | 540 | 1.54 | .93 |
| 4. I skipped items I was unable to answer | 364 | 2.29 | 1.43 | 542 | 2.55 | 1.35 |
| Time management | | | | | | |
| 5. Occasionally, I got stuck and spent a lot of time on difficult questions | 363 | 2.95 | 1.24 | 542 | 2.90 | 1.20 |
| 6. Ensuring accuracy was more important than ensuring time to answer all items | 361 | 2.48 | 1.18 | 542 | 2.22 | 1.04 |
| 7. I use the full time allowed for the test | 367 | 3.96 | 1.24 | 536 | 3.97 | 1.13 |
| 8. I kept track of time during the whole test | 363 | 3.46 | 1.40 | 541 | 3.41 | 1.32 |
| Test-wiseness, if not sure about the answer | | | | | | |
| 9. I look for clues in the item stem and the response options | 361 | 3.47 | 1.19 | 542 | 3.46 | 1.09 |
| 10. I use elimination processes | 363 | 3.67 | 1.18 | 541 | 3.80 | 1.01 |
| 11. Guessing or random choice of response option | 363 | 3.64 | 1.29 | 540 | 2.99 | 1.36 |
| 12. I like MC tests | 362 | 3.86 | 1.17 | 539 | 3.87 | 1.05 |

*p < .05.
**p < .01.
developed in agreement with earlier research in the area, although some items were developed specifically with the present test context in mind.

**Test anxiety and test-taking motivation**
Test anxiety was measured with nine items, asking primarily for emotional and cognitive aspects of test anxiety (fear of failing the test, worries about the difficulty of the test, whether the test situation as such made the test-taker feel stressed or nervous). The items were rated on a 5-point Likert scale from strongly disagree to strongly agree (internal consistency, coefficient α, for these items was .86, suggesting content homogeneity).

Perceived importance and motivation to spend effort on the test was measured with three items about whether the test-taker felt motivated to do his or her best and whether a good result was perceived as important. These items were also rated on a 5-point Likert-type scale ranging from strongly disagree to strongly agree (internal consistency, coefficient alpha for these questions was .76, also suggesting content homogeneity).

The anxiety and motivation scales were adapted from previous studies in different Swedish assessment contexts (Eklöf & Nyroos, 2013; Knekta & Eklöf, 2015), where they have demonstrated acceptable psychometric properties.

**Statistical analyses**
χ²-tests were used to examine group differences in use of the categorical variables (i.e. the test preparation strategies) and independent-sample t-tests were used to examine group differences in use of the continuous variables (i.e. during test strategies, motivation and test anxiety). As we have a large sample size and the variables are frequency count parametric t-tests were used. Analyses were first performed separately for high and low achievers and gender. Further, two-way, between-group analyses of variance (ANOVA) were used to examine the main effects of gender and achievement level, as well as the interaction effect between gender and achievement level (third section of the result). In the ANOVAs the result from Levene’s test of equality of error variance was inspected, and when the assumption of equal variance was violated the α level was set to .01. The α value was set to .05 for all other analyses. To estimate effect sizes we use the Phi coefficient (φ) for the χ²-tests, Cohen’s d for the independent-samples t-test, and partial eta squared (η²) for the ANOVAs.

**Results**

**Differences in reported test-taking behaviour between high and low achievers**

**Test-preparation**
The use of test preparation strategies among high and low achievers has been examined using χ²-tests as the items were assessed on a yes-or-no scale (Table 1). The results show that there was a significant association between achievement level and whether or not they used test-preparation strategy five (χ²(1) = 4.01, p = .045, φ = .07) and six (χ²(1) = 15.8, p < .001, φ = .13). Thus, high achievers reported to a higher extent that they have educated themselves about the test and its content (strategy six), while low achievers to a higher extent reported that they viewed old tests without actually practicing on them (strategy five). The differences between the groups in terms of effect sizes were small, however. The result did not show any differences between the groups in use of the other test-preparation
strategies. When only the groups categorised as extreme high/low achievers were included in the analysis, the result remains for strategy five \(\chi^2(1) = 7.57, p = .006, \phi = .12\), and six \(\chi^2(1) = 18.83, p < .001, \phi = .20\), with slightly larger effects sizes. There were no significant differences between the extreme groups for the other test-preparation strategies.

**During-test strategies**

The use of during-test strategies, including structural organisation, time management and test-wiseness, has been examined using independent samples \(t\)-tests.

For the test-taking strategies categorised as structural organisation (strategy one to four in Table 2) the \(t\)-tests indicated statistically significant differences between high and low achievers in three of the four strategies (strategy one, three and four). Low achievers reported to a higher extent that they looked through the test before they started \((t(686.6) = 3.45, p = .001, d = .23)\), and that they jumped between items and different subtests \((t(703.4) = 2.42, p = .016, d = .16)\); see Table 2 for \(M\) and SD). Further, high achievers reported that they skipped items they did not know the answer to a higher extent compared to low achievers \((t(904) = −2.77, p = .006, d = .19)\). Again, the effect sizes were small. When only including the extreme low and high achievers in the analyses the results again show statistically significant differences for strategy one \((t(278.9) = 3.07, p = .002, d = .31)\), three \((t(281.1) = 3.06, p = .002, d = .30)\), and four \((t(475) = −3.43, p = .001, d = .33)\), and with larger effect sizes. However, in this analysis, the use of strategy two also is significantly different between the two groups. \((t(479) = 2.10, p = .03, d = .21)\) Thus, extreme low achievers reported to a higher extent that they answered the items in the order they were displayed in the test booklet \((M = 4.2, SD = 1.1)\) compared to extreme high achievers \((M = 3.9, SD = 1.2)\).

The results from the test-taking strategies categorised as time management (strategy five to eight in Table 2) show that there are statistically significant differences between low and high achievers in one of the four strategies (six). Low achievers to a higher extent reported that it was more important to ensure accuracy in their answers than certifying time to answer all items \((t(700.2) = 3.36, p = .001, d = .23)\); see Table 2 for \(M\) and SD). When only comparing the extreme low and high achievers there is still a significant difference in strategy six \((t(283.7) = 2.89, p = .002, d = .28)\), and the effect size is somewhat larger.

In the test-wiseness category (strategy 9–12 in Table 2), the results show statistically significant differences between high and low achievers in one of the four strategies. Low achievers to a higher extent reported guessing (or randomly choosing a response option) if they did not know the answer to an item \((t(901) = 7.20, p < .001, d = .49)\); see Table 2 for \(M\) and SD). When only including the extreme high and low achievers the results remains, with large effect size, for guessing \((t(474) = 7.33, p < .001, d = .72)\). In addition, with only the extremes included, strategy 10 also differed between the groups \((t(267.2) = −2.15, p = .032, d = .22)\) but the effect size is small. Thus, extreme high achievers to a higher extent reported that they used elimination processes \((M = 3.76, SD = 0.99)\) compared to extreme low achievers \((M = 3.51, SD = 1.26)\).

**Test anxiety and motivation among high and low achievers**

To examine whether the level of test anxiety and motivation differs between high and low achievers, an independent samples \(t\)-test was performed. The results show, not surprisingly, that there is a statistically significant difference between the groups in test anxiety \((t(875) = 6.43, p < .001, d = .44)\). The low achievers reported a higher degree of test-anxiety.
(M = 30.2, SD = 8.1) compared to the high achievers (M = 26.6, SD = 7.9). The results also show a statistically significant difference between level of reported motivation between the two groups (t (712.3) = −3.04, p = .002, d = .21), a small effect size though. High achievers reported a slightly higher degree of motivation (M = 11.9, SD = 2.5) compared to low achievers (M = 11.3, SD = 2.8). When only including the extreme low and high achievers the result still remains for test anxiety (t (460) = 8.89, p < .001, d = .86), large effect size, and motivation (t (273.2) = −2.25, p = .025, d = .22).

**Differences in reported test-taking behaviour between females and males**

**Test-preparation**
To examine differences between males and females in use of test preparation strategies Chi-2 tests were performed. The results show that there was no significant association between group and whether or not they used specific test-preparation strategies (Table 3).

**During-test strategies**
To examine whether females and males differ in their use of during-test strategies independent-sample t-test were performed.

The result show no statistically significant differences between males and females in use of the test-strategies categorised as structural organisation (see Table 4). In the test-taking strategies categorised as time management the results show a statistically significant difference between males and females in three (six, seven and eight) of the strategies. Males to a significantly higher extent reported that it was more important to ensure accuracy in their responses than certifying time to answer all items (t (1099) = 2.23, p = .026, d = .14; see Table 4 for M and SD). There was also a significant difference between the groups in time used for the test (t (769.25) = −2.81, p = .005, d = .17). Females reported to a higher extent to have used all time allowed for the test, and to have kept track of time (t (1102) = −5.02, p = .000, d = .32).

In the test-wiseness category the results show statistically significant differences between males and females in one of these strategies (strategy eleven). Females reported a significantly higher extent in use of guessing or random choice of response option (t (1100) = −6.83, p < .001, d = .42).

**Test anxiety and motivation among females and males**
To examine if the level of test-anxiety and motivation (and importance) among females and males differs an independent samples t-test were performed. The result shows that

<table>
<thead>
<tr>
<th>Test-preparation strategies</th>
<th>Females</th>
<th></th>
<th>Males</th>
<th></th>
<th>All</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Use (n)</td>
<td>%</td>
<td></td>
<td>Use (n)</td>
<td>%</td>
<td>Use (n)</td>
<td>%</td>
</tr>
<tr>
<td>1. Participated in a course or workshop</td>
<td>15</td>
<td>2.1</td>
<td>15</td>
<td>3.6</td>
<td>30</td>
<td>2.7</td>
</tr>
<tr>
<td>2. Studied the areas included in the test</td>
<td>240</td>
<td>33.7</td>
<td>131</td>
<td>31.5</td>
<td>371</td>
<td>32.9</td>
</tr>
<tr>
<td>3. Practice – old tests with time limits</td>
<td>178</td>
<td>25.0</td>
<td>101</td>
<td>24.3</td>
<td>279</td>
<td>24.7</td>
</tr>
<tr>
<td>4. Practice – old tests without time limits</td>
<td>259</td>
<td>36.3</td>
<td>158</td>
<td>38.0</td>
<td>417</td>
<td>36.9</td>
</tr>
<tr>
<td>5. Just viewing old tests to get an idea of the content</td>
<td>377</td>
<td>52.9</td>
<td>214</td>
<td>51.4</td>
<td>591</td>
<td>52.3</td>
</tr>
<tr>
<td>6. Information about the test and its content areas</td>
<td>468</td>
<td>65.6</td>
<td>261</td>
<td>62.7</td>
<td>729</td>
<td>64.6</td>
</tr>
</tbody>
</table>
there is a statistically significant difference in reported test anxiety among females and males ($t(1071) = -6.33, p < .001, \phi = .40$). Females reported a higher degree of test-anxiety ($M = 29.4, SD = 8.2$) compared to males ($M = 26.1, SD = 8.1$). No significant difference in reported motivation was found.

**Gender differences within achievement level groups**

In the two previous sections we have seen that high and low achievers, as well as females and males, differ on some aspects of test-taking behaviour. This section reports whether the significant differences found between low and high achievers in use of test-taking strategies, as well as perceived motivation and test anxiety, are the same for females and males.

In the reported use of test-preparation strategies, strategy five and six were found to be significantly different between high and low achievers. Separate $\chi^2$-tests were made for females and males to examine whether the results vary. The results show that there was a significant difference in strategy six between high and low achieving males ($\chi^2(1) = 15.18, p < .001, \phi = .21$), and females ($\chi^2(1) = 4.95, p = .021, \phi = .09$). High achieving males and females to a higher extent reported that they have educated themselves about the test and its content. Regarding strategy five there was no statistically significant difference found between high and low achievers regardless of gender. Thus, the differences found between high and low achievers in the preparation strategies do not vary between genders.

To examine whether the observed differences between the achievement groups in during-test strategies are the same for males and females, two-way, between-groups, ANOVA were performed. In these results, only possible main effects of gender are important as we already have presented differences related to achievement level.

In the category *organisational structure*, significant differences were found between low and high achievers in three (one, three and four) of the four strategies. The result from the

---

**Table 4. Frequency (N), mean (M) and standard deviation (SD) of the during-test strategies among females and males.**

<table>
<thead>
<tr>
<th>During-test strategies</th>
<th>Females</th>
<th></th>
<th></th>
<th>Males</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>M</td>
<td>SD</td>
<td>N</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Structural organisation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. I looked through the test before I started</td>
<td>700</td>
<td>1.75</td>
<td>1.15</td>
<td>406</td>
<td>1.64</td>
<td>1.07</td>
</tr>
<tr>
<td>2. I answered the items in the order they were displayed in the test booklet</td>
<td>700</td>
<td>3.98</td>
<td>1.24</td>
<td>409</td>
<td>4.11</td>
<td>1.19</td>
</tr>
<tr>
<td>3. I jumped between items and subtests</td>
<td>691</td>
<td>1.64</td>
<td>.10</td>
<td>406</td>
<td>1.59</td>
<td>.10</td>
</tr>
<tr>
<td>4. I skipped items I was unable to answer</td>
<td>699</td>
<td>2.46</td>
<td>1.42</td>
<td>405</td>
<td>2.43</td>
<td>1.35</td>
</tr>
<tr>
<td>Time management</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Occasionally, I got stuck and spent a lot of time on difficult questions</td>
<td>699</td>
<td>2.91</td>
<td>1.23</td>
<td>404</td>
<td>3.03</td>
<td>1.22</td>
</tr>
<tr>
<td>6. Ensuring accuracy was more important than ensuring time to answer all items*</td>
<td>696</td>
<td>2.28</td>
<td>1.10</td>
<td>405</td>
<td>2.44</td>
<td>1.13</td>
</tr>
<tr>
<td>7. I use the full time allowed for the test**</td>
<td>701</td>
<td>4.10</td>
<td>1.09</td>
<td>407</td>
<td>3.90</td>
<td>1.23</td>
</tr>
<tr>
<td>8. I kept track of time during the whole test**</td>
<td>697</td>
<td>3.58</td>
<td>1.33</td>
<td>407</td>
<td>3.15</td>
<td>1.37</td>
</tr>
<tr>
<td>Test-wiseness, if not sure about the answer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. I look for clues in the item stem and the response options</td>
<td>697</td>
<td>3.43</td>
<td>1.15</td>
<td>406</td>
<td>3.52</td>
<td>1.08</td>
</tr>
<tr>
<td>10. I use elimination processes</td>
<td>699</td>
<td>3.81</td>
<td>1.07</td>
<td>406</td>
<td>3.70</td>
<td>1.07</td>
</tr>
<tr>
<td>11. Guessing or random choice of response option**</td>
<td>697</td>
<td>3.54</td>
<td>1.32</td>
<td>405</td>
<td>2.99</td>
<td>1.36</td>
</tr>
<tr>
<td>12. I like MC tests</td>
<td>698</td>
<td>3.89</td>
<td>1.12</td>
<td>404</td>
<td>3.87</td>
<td>1.05</td>
</tr>
</tbody>
</table>

*p < .05.

**p < .001.
ANOVAs showed that there is no interaction effect between achievement level and gender on reported use of strategy one \((F (1, 901) = 1.043, \text{MSE} = 1.22, p = .308)\), three \((F (1, 894) = 0.51, \text{MSE} = 0.97, p = .822)\) and four \((F (1, 902) = 3.05, \text{MSE} = 1.92, p = .081)\). Also, no significant main effect for gender was found. This means that males and females do not differ in use of these strategies.

Regarding time management, one strategy (six) was found to be significantly different between high and low achievers. The results from the ANOVAs showed no interaction effect between achievement level and gender in use of strategy six \((F (1, 899) = 0.101, \text{MSE} = 1.20, p = .751)\). However, there was a main effect for achievement level \((F (1, 899) = 14.25, \text{MSE} = 1.20, p < .001, \eta_p^2 = .016)\), as well as for gender \((F (1, 899) = 8.21, \text{MSE} = 1.20, p = .004, \eta_p^2 = .009)\), but the effect sizes are small. Males reported to a significantly higher extent (low achievers \(M = 2.7, \text{SD} = 1.1\); high achievers \(M = 2.3, \text{SD} = 1.1\) ) that it was more important to ensure accuracy than to ensure time to answer all questions compared females (low achievers \(M = 2.4, \text{SD} = 1.2\); high achievers \(M = 2.1, \text{SD} = 1.0\) ), regardless of achievement level.

In the strategies categorised as test-wiseness only one strategy (11) was found to be different between low and high achievers. The result from the ANOVA showed no significant interaction effect between gender and achievement level in use of this strategy \((F (1, 899) = 0.003, \text{MSE} = 1.74, p = .954)\). There was a significant main effect for achievement level \((F (1, 899) = 36.87, \text{MSE} = 1.74, p < .001, \eta_p^2 = .039)\), as well as for gender \((F (1, 899) = 22.52, \text{MSE} = 1.74, p < .001, \eta_p^2 = .024)\). Regardless of achievement level, females to a significantly higher extent (low achievers \(M = 3.8, \text{SD} = 1.2\); high achievers \(M = 3.2, \text{SD} = 1.3\) ) than males (low achievers \(M = 3.3, \text{SD} = 1.3\); high achievers \(M = 2.7, \text{SD} = 1.4\) ) reported using random guessing.

Finally, the differences found between the achievement groups in test anxiety and motivation were examined. No statistical significant interaction effect was found between gender and achievement level in reported test anxiety \((F (1, 873) = .09, \text{MSE} = 62.52, p = .758)\). However, there was a significant main effect for achievement level \((F (1, 873) = 27.52, \text{MSE} = 62.52, p < .001, \eta_p^2 = .031)\), and for gender \((F (1, 873) = 25.86, \text{MSE} = 62.52, p < .001, \eta_p^2 = .029)\). Thus, females reported a significantly higher degree of test anxiety (low achievers \(M = 31.2, \text{SD} = 7.6\); high achievers \(M = 27.9, \text{SD} = 8.1\) ) than males (low achievers \(M = 28.0, \text{SD} = 8.7\); high achievers \(M = 25.1, \text{SD} = 7.5\) ) regardless of achievement level. As concerns reported motivation no significant interaction effect was found \((F (1, 899) = 1.09, \text{MSE} = 6.98, p = .296)\). There was a significant main effect for achievement level \((F (1, 899) = 6.31, \text{MSE} = 6.98, p = .012, \eta_p^2 = .007)\), but not for gender \((F (1, 899) = 0.61, \text{MSE} = 6.98, p = .435)\). This means that males and females do not differ in motivation.

**Discussion**

The result of the present study showed that there are some differences between high and low achievers, and between females and males, when it comes to reported use of test-taking strategies. It was also shown that low achievers reported higher levels of test anxiety, and that females reported higher levels of test anxiety regardless of achievement level when compared to males. Further, high achievers reported somewhat higher levels of motivation regardless of gender.
The first area of test-taking strategies we examined was test-preparation strategies. To practice on test items similar to those in the test is considered to be one of the most effective preparation strategies (Bond & Harman, 1994). Surprisingly, the results suggest that both high and low achievers practice on old test forms to the same extent, at least when they take the test for the first time. Further, a closer look at the proportion practicing on old test with, or without time limits, shows that the number of test-takers using this strategy is rather low in both groups (see Table 1). It is, however, possible that test-wise test takers tend to practice old test forms, with or without time limits, as an effective after-test strategy instead of as a preparation strategy (Dodeen, 2008). Additional results are that low achievers tend to view old tests to get an idea of the content without actually practice on them to a higher degree than high achievers, and that high achievers seem to educate themselves about the test and its content to a higher extent compared to low achievers. These results are in line with earlier studies suggesting that high achievers tend to prepare better (Hong et al., 2006), and that preparation has a positive effect on test performance (Bicak, 2013).

The second area examined was the test-takers reported during-test strategies. We found differences between high and low achievers in three of the four strategies categorised as structural organisation. It appears as if low achievers looked through the whole test before they began to a higher extent than high achievers. This strategy is known to be an ineffective test-taking strategy (Ellis & Ryan, 2003), probably because this strategy is time consuming, and therefore the test-taker will end up with less time to actually deal with the items and they may even fall short of completing the entire test. Another explanation to why this is an ineffective strategy could be that if the test-taker finds the test difficult this might lead to higher levels of test anxiety or worry which in turn could have a negative impact on the performance (see, e.g. Naylor, 1997). The results also showed that high achievers seem to use the more time effective strategy to skip difficult items. This result is in line with earlier research comparing groups of different achievement levels (Kim & Goetz, 1993; McClain, 1983). Low achievers also jumped between items and subparts of the test to a higher extent than high achievers, and this is probably mainly an effect of not being able to answer many of the questions.

Concerning during-test strategies related to time, time management, differences between genders, as well as achievement levels, were found. It seems as if low achievers find it more important to ensure accuracy than time to answer all questions. When examining this further, the results show that males reported this to be more important, regardless of achievement level, when compared to females. Further, there was no difference between low and high achievers in whether they used all time allowed for the test or whether they kept track of time. This result is somewhat surprising as these strategies are regarded as effective (Ellis & Ryan, 2003). However, when comparing females and males the results showed that females reported keeping track of time and using all timed allowed to a significantly higher extent than males. It is difficult to know why strategies related to time are more used by females, but these results suggests that females use effective time management strategies to a higher extent than males, regardless of achievement level.

In the analysis of the strategies categorised as test-wiseness only the use of random guessing differed significantly between achievement levels. Low achievers, not surprisingly, reported to use random guessing to a higher degree than high achievers. Even though educated guessing is considered to be an effective strategy, to randomly choose an answer option is not (Bond & Harman, 1994). However, maybe more surprisingly, when examining
this result further it was found that females used this strategy to a higher degree compared to males, regardless of achievement level. This result is not in line with earlier research showing that women are more reluctant to guess with or without penalty for wrong answers (Baldi, 2014; Ben-Shakhar & Sinai, 1991; Hirschfeld et al., 1995). Further, extreme high achievers seem to use elimination processes to a higher degree when compared to extreme low achievers. This result has also been found in earlier studies comparing test-taking strategies among high and low achievers (see, e.g. Hong et al., 2006).

The third area examined in the present study was motivational and emotional concerns. The analyses of the group differences in terms of test anxiety provided quite strong evidence that low achievers, especially females and extreme low achievers, tend to have higher levels of test anxiety. This result is in line with earlier research about test anxiety (Zeidner, 1998). Possible explanations to this result are that (i) it is related to an inability to suppress anxiety with decreased performance as a consequence or (ii) it is related to perceived difficulty as a source of anxiety, where perceived difficulty in turn is a consequence of low ability (Cassady & Johnson, 2002). The analysis of differences in terms of motivation showed that low achievers tend to have slightly lower levels of motivation, when compared to high achievers, regardless of gender. The perceived importance of the SweSAT should be about the same for all test-takers as it is a high-stakes test situation; however, it is of course possible that their motivation varies with regards to intended use of the test (Cheng et al., 2014). Considering it is well known that it is beneficial to have practiced the SweSAT in a real setting, a possible explanation might be that the test scores are only to be used as an opportunity to practice the test situation.

In sum, the results indicate that there are some differences in reported test-taking behaviour between high and low achievers. Not only do high achievers tend to use more successful test-taking strategies, they also report less test anxiety and slightly higher levels of test motivation. Further, the results indicate differences in test-taking behaviour between females and males. Females seem to use random guessing to a larger extent, but they also appear to be more concerned about time, and report higher levels of test anxiety. (It should be noted that the size of the differences were in many cases small.) These differences are larger (in terms of effect sizes) when only including the extreme high and low achievers, suggesting more solid results even though the degrees of freedom are reduced. However, the differences found between high and low achievers could, to some degree, be due to age rather than performance on the SweSAT. Some studies suggest that there might be a maturity effect on test-wiseness skills at the college level (Geiger, 1997), and further research should be made to investigate this possibility. From the present study design we also cannot make any conclusions about whether differences in performance are actually due to differences in test-taking behaviour, but as high-performing individuals reported somewhat different test-taking behaviour than low performers, and previous studies have shown that successful test-taking strategies are related to a better test performance, this is also an important area for continued studies. This was the first administration of these items in a Swedish assessment context, and it is important to further refine the instrument to attain a basis for more valid conclusions about test-taking behaviour.

Still, the results in the present study indicate that the SweSAT might measure something more than just pure knowledge of the content being assessed. However, whether test-taking behaviour brings construct-irrelevant variance to test scores or not depends on the definition of the construct. As Sternberg (1998) claims that this type of expertise is required to be
successful in school and in many work situations, and a purpose of the SweSAT is to predict success in higher education, one could argue that the potential contribution of test-taking strategies on the test scores is not irrelevant. Sternberg’s view is supported by Dodeen et al. (2014) who claims that they are transferable, i.e. possible to use across many different settings and conditions, and moreover, that ‘test-taking skills are useful in a students’ practical life, where they may benefit their effective use of time, ability to set priorities, ability to work both fast and accurately, and to make sure ideas become directly evident’ (p. 1). It has also been found that possession of these skills might reduce test-anxiety and enhance motivation (Dodeen, 2009). Hence, the results in this study indicate educational implications for stakeholders and test developers. It seems clear that instructions and information about effective test-taking behaviour, before, during and also after the SweSAT (and other tests), need to be implemented. Not only to avoid biased test scores, but also because this type of skills may be important to be a successful test-taker and student.

Disclosure statement
No potential conflict of interest was reported by the authors.

Funding
This research was funded by the Swedish Research Council [grant number 2012-5075].

Notes on contributors
Tova Stenlund is assistant professor at the Department of Psychology, Umeå University, Sweden. She has a PhD degree in educational measurement and her research interests lie in the fields of educational psychology and cognitive psychology, with a particular focus on validity in test and assessment results, test-taking behaviour, and effects of repeated testing on memory and learning.

Hanna Eklöf is associate professor at the Department of Applied Educational Science, Umeå University, Sweden. She has a PhD degree in educational measurement and her research interests lie in the fields of educational psychology and large-scale achievement testing, with a particular focus on the test-takers’ perspectives in assessment contexts such as TIMSS, PISA, and national tests.

Per-Erik Lyrén is assistant professor at the Department of Applied Educational Science, Umeå University, Sweden. He has a PhD degree in educational measurement and his research interests lie in the fields of college admissions testing and measurement of patient-reported outcomes, with a particular focus on score reporting, validation, equating, responsiveness and test-takers’ perspectives on testing.

References


