Just Enough Time to Level the Playing Field: Time Adaptation in a College Admission Test

Marcus S. Hjärne

Department of Applied Educational Science, Umeå University, Umeå, Sweden

ABSTRACT

Extended time is a commonly-used test adaptation for standardised high-stakes tests. In this study, extended time provided for test-takers with dyslexia is examined. Data from standard versions of the Swedish Scholastic Aptitude Test (SweSAT) and data from test administrations where extra time is provided was used. Indications are that the standard version of the test is speeded and that the extra time provided improves test results for test-takers with dyslexia. There are, however, no conclusive indications supporting differential speededness due to the extra time. Findings from this study shows the beneficial effects of providing extra time for test-takers with dyslexia, but it does not conclusively disprove that this extra time cause an unfair advantage.

ARTICLE HISTORY

Received 9 July 2019
Accepted 13 April 2020

KEYWORDS

Adaptation; accommodation; fairness; bias; extended time

Introduction

Standardised tests are used to provide fundamental and significant information on individuals and groups. One aspect of standardisation of tests is to ensure that all test-takers experience the same testing condition. It is essential for the validity of test score interpretation and use that the test score is influenced by the test construct only. The test construct is the attribute that is postulated to be reflected by the test performance (Cronbach & Meehl, 1955). For a more elaborated description of how to validate the test construct see Messick (1989). Standardisation may cause unfair procedures in test administration by introducing construct-irrelevant barriers for some test-takers. For example, written texts provided for visually-impaired test-takers, or exhaustive testing procedures for individuals with narcolepsy. In order to overcome such construct-irrelevant barriers, adaptations to standardised tests are usually introduced.

Extra time is one of the most commonly used adaptation of standardised tests (Sireci et al., 2005). Extra time is supposed to help test-takers who, for some reason, need more time to complete all the items and therefore the test. The assumption is that the extra time will help the test-taker to overcome a construct-irrelevant barrier when speed is not included in the construct. However, providing extra time is an exception from the standardisation that ensures equal terms in the comparison of test results. In general, such alternations require the reestablishment of the argument used to validate the standardised test (Kane, 2013; Sheinker et al., 2004).

Data from the Swedish Scholastic Aptitude Test (SweSAT) was used in this study. The SweSAT is a high-stakes, multiple choice test used for college admission in Sweden. Time (or speed) is not an intended component of the SweSAT test construct. The assumption when providing extra time is that while the time is adequate for others, extra time will help test-takers with dyslexia to overcome
a construct-irrelevant barrier (lack of time). In recent years, an increasing number of test-takers have been provided with extra time on the SweSAT. The group of test-takers with diagnosed dyslexia is by far the largest group of test-takers provided with extra time and this group is growing every year (an increase of nearly 100 percent between 2011 and 2019). Literature on the effect of extra time is inconsistent as to whether extra time is a fair alteration of high-stakes standardised admission tests and how findings can improve the validity of such test score use. Existing research on the effect of extra time typically uses an experimental design that does not resemble a high-stakes admission test setting.

The aim of this study was to examine whether providing extra time is a fair adaptation that helps test-takers with dyslexia\(^1\) to overcome construct-irrelevant barriers when taking a college admission test. The hypothesis is that providing extra time can cause test bias. Is the standard version of the test speeded? Does extended time help test-takers with dyslexia to overcome construct-irrelevant barriers? Are there any indications of the test being more or less speeded for test-takers provided with extra time due to their dyslexia?

**Background and Previous Research**

**Test Adaptations**

A generic three-tier structure of categorisation of different types of test adaptations has been presented by Sheinker et al. (2004). The proposed structure categorises adaptation on the basis of the “(...) likelihood of changing the construct or target skill/knowledge of a test” (Kettler, 2012, p. 57). The first category consists of accommodations that are far from affecting the test construct. Examples of accommodations in this category include the use of large prints and changes in testing location (Sheinker et al., 2004). The second category of adaptations may affect test results and the interpretation based on these results. Extended time is an example of the second category of test adaptation (ibid.). The validity argument when introducing extra time rests upon the belief that the adaptation will eliminate barriers that establish construct-irrelevant variance without altering the test construct (Kettler, 2012). The third category of accommodations is based on adaptations that relate directly to changes in the test construct. Such adaptations are often considered to be modifications of the test. Accommodations that give all test-takers equal opportunities can be considered unproblematic, while it may be considered to cause unfair advantages for the students who take an adapted version of the test if the adaptation benefits every test-taker and not only test-takers that needs an accommodated version of the test.

The 2014 standards of the American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] (2014), classify different test adaptations as either test accommodations or test modifications. Adaptations are made to increase accessibility to the test for individuals who are otherwise facing construct-irrelevant barriers and include changes in test content, test format and administration conditions. Accommodations are defined as adaptations that do not change the test construct while modifications change the test construct and, as a consequence, the score meaning. For this study the categorisation provided in Table 1 will be used. This categorisation is coherent with the categorisation by Sheinker et al. (2004) and is a development of the 2014 standards to include the second category of test accommodations.

**The Maximum Potential Thesis**

Current theories used to frame construct-irrelevant group differences when an adaptation of a standardised test is provided include the interaction hypothesis and differential boost. In the perspective

---

\(^1\)This group is defined by ICD-10 code f81.0 and f81.3 and will be referred to as the dyslectic group in the manuscript.
of adaptation of the test using extended time, the rational for both the interaction hypothesis and differential boost can be explained by the ideas of the maximum potential thesis. The maximum potential thesis postulates that test-takers without the need of any adaptation would not benefit from extended time since their maximum potential is shown under timed conditions (Cawthon et al., 2009; Zuriiff, 2000), and as Zuriiff (2000) states that “(...) many students with learning disabilities process information more slowly and do not operate at full potential unless given the extra time.” (p. 101).

Differential boost is described and used in many studies on the validity of test adaptations (e.g., Cawthon et al., 2009; Fuchs et al., 2005; Kettler, 2012; Lang et al., 2008). The primary idea of differential boost is that if the maximum potential thesis is valid, then an accommodation is expected to boost the performance of students with disabilities more than the population in general. The interaction hypothesis (Sireci et al., 2005) can be described as a version of differential boost. It postulates that the results of students with disabilities are helped by accommodations while other test-takers are not. An accommodation is expected to compensate for the barrier that their disabilities cause. Hence, the test result is an interaction between the accommodation and the student group. The absence of such interaction would show in equal benefits from the adaptation for all test-takers.

**Time Accommodations**

Some summaries of research on time accommodations have been published (e.g., Jennings, 2014; Sireci et al., 2005). Sireci et al. (2005) conclude that existing research is inconsistent regarding extended time supporting the interaction hypothesis. Sireci et al. (2005), Jennings (2014) presents a summary of research on the utility of extended time as adaptation and concludes that research in general does not suggest that time adaptations provide a differential boost and that all students, regardless of disability status, are equally benefitted from extra time. There are wide differences in the settings of the tests included in existing studies. Many important factors show differences between studies (e.g., test content, the stakes of the test, amount of extended time, length of test, age of test-takers, disability status of test-takers). Due to the conflicting research findings, recommendations for high-stakes testing programmes with accommodations should be to specifically study the effect of the accommodation provided.

Studies of extended time as an accommodation for students with disabilities typically uses a methodological design where both students with disabilities and students without disabilities are provided with extra time (e.g., Elliott & Marquart, 2004; Lewandowski et al., 2008, 2013; Mandinach et al., 2005). In this typical case of current research, different groups are given the treatment (the accommodation) and the effect of the treatment is used to inform on the validity of the treatment. This is an experimental design that can control the time factor and the benefits are self-evident. Nonetheless the typical experimental design often comes with some disadvantages. First, it is difficult and costly to obtain large enough samples when group sizes need to be quite extensive for calculations of reliable effect sizes. Second, the typical experimental setting has very little in common to the setting of a high-stakes admission test. Test motivation and anxiety are functions of the importance of the test results and an experimental setting will risk diminishing the effects and the effect on test results (Cheng et al., 2014). Third, validation of accommodation could be argued to be preferably a standard procedure in the ongoing test validation procedure and not a single experiment carried out in an
artificial environment. Using real test data could potentially discover ways of interpreting the effect of extended time in an ongoing validation process.

**Method**

**Data and Participants**

The SweSAT is administrated twice every year and is used to select among qualified applicants to higher education in Sweden. The test is based on classical test theory (CTT) and aims at measuring scholastic proficiency. A test score, $X$, in classical test theory is defined by: $X = T + E$, where $T$ is the true score and $E$ is a random error score. Subsequently, in an error free measure, the true score equals the observed score. The assumptions in the classical test model are that true scores and error scores are uncorrelated, the average error score in the population of examinees is zero, and error scores on parallel tests are uncorrelated (Hambleton & Jones, 1993).

The test is administrated in four separate booklets (and one extra pre-test booklet) with a break in between. It contains 160 items (plus 40 pre-test items) divided into eight sub-tests that form a verbal and a quantitative part, each containing four subtests. There is no penalty for wrong answers on the test and very few items (0.4%) are omitted by test-takers. The scoring is automated and anonymous, and the scores on each part are equated and normed on a 0.0 to 2.0 scale. In the test administration of spring 2018 (2018a), the mean score for the quantitative part was 0.91 with a standard deviation of 0.46 and for the verbal part the mean score was 0.85 with a standard deviation of 0.43.

This study uses a non-experimental design with real test data from several test administrations. Descriptive statistics from 2017b (autumn 2017) and 2018a (spring 2018) are shown in Table 2. The distribution of age and male/female test-takers are roughly the same in the group of test-takers with no accommodation as in the group of test-takers with extended time due to dyslexia. Test-takers with extended time due to dyslexia performed better than test-takers with no accommodation in total and on the quantitative part on both administrations (Table 2).

Several different analyses were conducted to answer the research questions. For each analysis, a subset of the data material was used. The subsets and the reason for using the specific part of the material are presented in connection with the respective analysis.

**Is the Standard Version of the Test Speeded?**

The 2018a data was used to examine potential test speededness in the standard version of the test. The 2018a test was administrated in several different versions containing the same items, but in different orders. For example, item one in version one was placed as item six in another version. Items were moved minimum zero and maximum eleven positions. No test-taker was given two

| Table 2. Descriptive statistics of test results for test-takers with time accommodation due to dyslexia and without accommodations. |
|---------------------------------------------------------------|
| Autumn 2017                                                   |
| No accommodation ($N = 45,776$)                               |
| Mean score total 0.88, Mean score quant 0.91, Mean score verb 0.84 | 62% 21% 9% 5% 2% 46% |
| SD 0.40, 0.46, 0.44                                           |
| Extended time ($N = 997$)                                    |
| Mean score total 0.89, Mean score quant 1.01, Mean score verb 0.76 | 58% 25% 9% 4% 1% 46% |
| SD 0.40, 0.48, 0.42                                           |
| Spring 2018                                                  |
| No accommodation ($N = 61,996$)                              |
| Mean score total 0.88, Mean score quant 0.91, Mean score verb 0.85 | 57% 25% 11% 6% 2% 49% |
| SD 0.39, 0.46, 0.43                                           |
| Extended time ($N = 1488$)                                   |
| Mean score total 0.90, Mean score quant 1.01, Mean score verb 0.78 | 54% 31% 10% 4% 1% 45% |
| SD 0.39, 0.49, 0.40                                           |
different versions of the same booklet. For administrative reasons, all test-takers provided with extra
time in spring 2018 took the same version of the 2018a test (Version 1). Three of the four booklets
were used in two different versions, Quantitative Booklet 1, Quantitative Booklet 2 and Verbal Book-
let 2 (Table 3).

Relationship between item position and item difficulty. Test scores can be based on different test
models. CTT and item response theory (IRT) are the two dominating models used for estimating
item parameters (and the examinees’ ability). The underlying assumptions of CTT are described
above. Following the estimation of the true score as a linear function of the observed score, the
difficulty of an item in CTT is defined as the proportion of correct answers on that item (sometimes
denoted as the p-value or the p-score of the item). The IRT approach can be based on a variety of
models for estimating the IRT parameters as a function of the observed score. Common assumptions
for the IRT estimation of item parameters are that (1) the items are unidimensional, (2) the items are
locally independent and (3) the examinees’ latent ability can be estimated independently of taking
the same test (Crocker & Algina, 2008). The first two assumptions are sometimes merged into
one (e.g., Hambleton & Jones, 1993).

One commonly-used IRT model is the one-parameter logistic model, sometimes called the Rasch
Model. Using this model, the probability $P_g$ of answering item $g$ correctly for examinees with the
latent ability $\theta$ is defined by:

$$P_g(\theta) = \frac{e^{(\theta - b_g)}}{1 + e^{(\theta - b_g)}} \quad \text{(1)}$$

where $b_g$ is the item difficulty.

Oshima (1994) found that the difficulty parameters of end-of-test items tend to be overestimated
on speeded tests. Consequently, a speeded test would potentially be recognised by an overestimation
of the difficulty parameters on later items.

In order to establish whether the SweSAT is speeded, the item difficulty estimated as the $b$-para-
meter in the IRT Rasch Model for parameter estimation was compared for the same item at two
different positions in the subtests. The $b$-parameter was estimated using a joint maximum likelihood
estimation (see. Cleary, 1968; Wright & Douglas, 1977) in the jMetrik software (Meyer, 2014) with
fixed mean $b = 0$ for each version. In accordance with the findings of Oshima (1994), the assumption
is that the IRT estimate of item difficulty will be higher when an item is positioned later in the test
where time becomes a factor. Hence, the correlations between change in item position and difficulty
were used as indicators of speededness.

The unidimensionality of the items used are theoretically founded in the test rational that specifies
two parts of four different subtests. Each of the subtest was analysed separately. Factor analysis sup-
ported the unidimensionality of the items analysed. The local independence of items might, to some

| Subset defined as test-takers taking | Normed total score | Normed quantitative score | Normed verbal score | n  |
|------------------------------------|-------------------|--------------------------|--------------------|----|
| Quantitative Booklet 1 Version 1   | Mean 0.87          | 0.90                     | 0.84               | 36,740 |
|                                    | SD 0.39            | 0.46                     |                    | 44  |
| Quantitative Booklet 1 Version 2   | Mean 0.88          | 0.92                     | 0.84               | 27,196 |
|                                    | SD 0.39            | 0.46                     |                    | 43  |
| Quantitative Booklet 2 Version 1   | Mean 0.85          | 0.89                     | 0.81               | 36,445 |
|                                    | SD 0.38            | 0.46                     |                    | 43  |
| Quantitative Booklet 2 Version 2   | Mean 0.90          | 0.93                     | 0.87               | 27,732 |
|                                    | SD 0.40            | 0.47                     |                    | 45  |
| Verbal Booklet 2 Version 1         | Mean 0.86          | 0.89                     | 0.83               | 36,107 |
|                                    | SD 0.39            | 0.46                     |                    | 43  |
| Verbal Booklet 2 Version 2         | Mean 0.88          | 0.92                     | 0.85               | 28,070 |
|                                    | SD 0.40            | 0.47                     |                    | 44  |

The scores are normed scores on the 0.00–2.00 scale.
extent, be violated since some items have common stimulus (e.g., multiple items in the same text in reading comprehension). All the items could be estimated using the Rasch Model, but very few items fit the model perfectly. The infit and outfit mean square values were in the range of 0.8–1.2, which is a suitable model fit for high-stakes tests (Bond & Fox, 2007).

**Does Extra Time Help Test-takers to Overcome Time as a Construct-Irrelevant Barrier?**

A positive effect of extended time for test-takers with dyslexia is expected if extra time helps to overcome time constraints as a construct-irrelevant barrier. In order to analyse the effect of extra time, two successive test administration sessions (2017b and spring 2018a) were used. The total number of test results was 110,258. Some test-takers took the test in two successive administrations and the total number of unique test-takers was 90,409, meaning that 19,849 test-takers took both tests. Only test-takers taking both tests were included in the analysis of the effect of extended time. In the group of test-takers taking both tests, 18,881 participants took both tests with no accommodation (NN Group), 469 were provided with extra time due to dyslexia at both administrations (AA Group), and 141 participants had no accommodation on the first and extended time as an accommodation due to dyslexia on the second administration (NA Group). Only five participants took the test with extra time on the first administration and without extra time on the second attempt. Due to few participants, this group was not used in the analyses.

The fact that a number of students with dyslexia took the test without time accommodations tells us that there is a number of unrecorded dyslexic test-takers in the group of test-takers with no accommodation. There should not be any test-takers without dyslexia in the group of test-takers with extended time since test-takers of SweSAT need to have a documented diagnosis issued to be considered as dyslectic and provided extra time.

The effect of extra time. It is known that the average test-taker increases their score due to a practice effect when retaking the SweSAT (Cliffordson, 2004). Previous research has shown score gains between 0.15 and 0.25 standard deviation units between the first and second sessions (ibid.).

The effect of extended time was analysed by a fixed effect model that enables to separate out any differences in repeated testing from the effect of adaptation (Murnane & Willett, 2010).

With a fixed effect, \( f_i \), dummies for administration (\( t = 0 \) for first administration an \( t = 1 \) for second administration) and group(s) (NA and NN) were regressed on the score, \( Y_{it} \). The AA-group serves as the reference group. This mean that the time variable captures the effect of repeated testing for that group and the other group parameters capture the difference between the respective group and the AA Group. In equation (2) the group variables are indexed with time. Meaning that \( \text{NA}_it \) and \( \text{NN}_it \) takes the value 1 for individuals included in each group, but only at administration 2.

\[
Y_{it} = a \times t + b \times \text{NA}_it + c \times \text{NN}_it + f_i + e_{it}
\]  

(2)

In order to estimate the effect of length without selection bias, random selection to the groups was assumed. However, there may possibly be a not entirely random distribution based on non-negligible factors for the effect of repeated testing. For example, it is possible that socio-economic background affects both which group you belong to and the effect of repeated sampling.

Findings from similar quasi-experimental analyses on the American SAT and ACT indicate three to four time larger score gains for re-testers with extended time when comparing to re-testers taking the test without extended time at both administrations (Camara et al., 1998; Ziomek & Andrews, 1998).

**Does Extra Time Cause Differences in Speededness?**

Data from the Quantitative and Verbal Version 1 of the test administration 2018a was used to analyse omitted items and differential item functioning (DIF) as indicators of group differences in
speededness. This version of the test was taken by all test-takers provided with any test adaptation and approximately 1/7 of the test-takers with no test adaptation. Thus, only the results from test-takers provided with equal position of test items in every booklet were included in the analysis. Indications of differential speededness was shown from the initial analyses of omitted items, and five different test administrations between 2015 and 2019 were used to verify the results (including over 350,000 tests). The additional datasets have approximately the same composition as the administration 2018a, described above.

**Group Differences in the Relationship Between Item Position and Indicators of Speededness**

Various methods can be used for analysing whether a test is differentially speeded. Many studies in the area have been published as research reports for test institutes (e.g., Mandinach et al., 2005; Schmitt et al., 1991; Secolsky, 1989). Traditionally, items not reached are used as an indicator of test speededness when a single test administration is analysed (e.g., Gulliksen, 1950; Stafford, 1971). However, it is most likely that test-takers at a high-stakes test with no penalty for wrong answers at least make a guess about the items at the end as the time runs out (approximately 99.6% answers each item on average in the SweSAT). Hence, this traditional approach was not used alone in this study. In addition, group-dependent speededness was proposed to be indicated by a Mantel-Haenszel approach for indicating differential item functioning, DIF (Mantel & Haenszel, 1959). The method for analysing group differences using each approach is described in more detail below and the two indicators (omitted items and DIF) are described under separate subtitles. But first some critical assumptions:

**Assumptions.** Both approaches are based on the assumptions of classical test theory (described earlier) and on the assumption that items at the end of each booklet are more influenced by test speededness. Another critical assumption for the proposed approaches is that the average test-taker takes the test items in the order that they are presented in each booklet. Stenlund et al. (2017) present result that strongly support this assumption for the SweSAT. There may also be some mediating factor involved that lead to misinterpretation of results using the suggested methods. Gender, age or general proficiency level can affect how the average test-taker behaves in skipping test items, taking chances, managing time etc. For example, 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). If the dyslectic test-takers provided with extra time deviates from the population in gender or age distribution, or in general proficiency, this must be accounted for. Group differences in age and gender distribution were checked and the groups were considered to be fairly equal.

**Omitted items indicator.** Using this approach, each booklet was analysed for omitted items. The assumption of this analyses was that more items will be omitted at the end of the booklet if the test is speeded and group differences in the amount of omitted items at the end of the test (each booklet) can be explained by differential speededness. The proportion of items omitted were compared between the group taking the test with extra time (ext) due to dyslexia and the group without any adaptation (std) of the test.

**Differential item function indicator.** Differences in speededness for the two groups compared can potentially be indicated by DIF. The assumption is that if the groups are similar then DIF would occur only if the time becomes a differentiating factor. DIF is usually examined between a focus group and a reference group. One of the groups, usually the reference group, is supposed to have an advantage or represents the norm for the population. DIF is shown when there are differences in how an item works when the compared groups are matched for performance (Dorans & Holland, 1992). Such differences occur when an item measures something other than underlying ability. For example, when time forms a hindrance to considering the correct answer. For this reason, DIF is commonly used as an indicator of construct-irrelevant group differences for individual items. The
main factor that defines the groups is the extra time provided for test-takers with dyslexia. Hence, differences in speededness is assumed to be indicated as DIF on test items that is influenced by speededness. DIF on items early in the test should not occur due to this assumption.

The procedure presented by Mantel and Haenszel (1959), and proposed as a method for detecting uniform DIF by Holland and Thayer (1988), was used for this approach. The method uses a focal group (test-takers with extended time due to dyslexia) and a reference group (test-takers with no accommodation) and compares the groups’ performances on each test item. The groups were matched on ability level, in this case the normed quantitative and verbal scores respectively. DIF on each item is shown if the null hypothesis, that there is no difference in item performance between the focal and the reference groups when controlling for ability, is rejected. The effect size was computed for each item (including no significant items) using the logarithmic measure, ln $\hat{\alpha}$, described in Wiberg (2009). The scale described by Zieky (1993) and modified by Wiberg (2009) was used for categorisation of effect sizes. A positive ln $\hat{\alpha}$ indicates that the item favours test-takers without accommodation and a negative ln $\hat{\alpha}$ indicates that the item favours test-takers with extended time due to dyslexia.

The categorisation of DIF used (ibid.) is as follows: A – Negligible DIF, ln $\hat{\alpha}$ is not significantly different from zero at the 0.05 level or $|\ln \hat{\alpha}| < 0.425$. B – Intermediate DIF $0.425 < |\ln \hat{\alpha}| < 0.638$. C – Large DIF, $|\ln \hat{\alpha}| \geq 0.638$.

The main focus in this study is not to identify individual items with potential bias due to DIF, instead ln $\hat{\alpha}$ was compared to the item position for each subtest and booklet.

**Comparing groups on the speededness indicators.** Correlations ($r_{\text{std}}$ and $r_{\text{ext}}$) between item position in each booklet and the two indicators of speededness (omitted items and DIF) for each group were calculated. In order to compare differences in correlations, z-scores were obtained by first transforming the correlation coefficients $r$ for the two groups using the formula presented by Fisher (1921):

$$r' = 0.5 \log_e \frac{1 + r}{1 - r}$$

Using this transformation, $r'$ is approximately normally distributed around the transformed population correlation, $\rho'$ (Howell, 2009). With additional information of group size, the standard error can be derived. The hypothesis that $\rho_{\text{ext}} - \rho_{\text{std}} = 0$ is equivalent to the null hypothesis $H_0: r'_{\text{ext}} - r'_{\text{std}} = 0$ and was tested by computing the normal curve deviate (Cohen et al., 2014):

$$z = \frac{r'_1 - r'_2}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}}$$

### Results

**Indication of Speededness in the Standard Version of the SweSAT**

The estimated difficulty using the Rasch $b$-parameter correlated significantly ($p < 0.001$) with the change in item position for all three booklets. The correlation coefficients can be considered to be large for the quantitative booklets ($r = 0.57$ and 0.55) and medium for the verbal booklet ($r = 0.48$) (Cohen, 1988). The correlations indicate that the standard version of the test was speeded for all test-takers taking the test without extra time (i.e., the standard version of the test).

**Improvement in Results due to Extended Time**

The normed score on the first and second test administration for each group is shown in Figure 1. The AA Group had slightly higher total test result than the NN Group at both test administrations
and the NA Group showed slightly lower total test results on the second test administration in relation to groups compared (Figure 1).

As expected, all three groups show improvement in test results between test sessions (Figure 1). Relatively large improvements in results were demonstrated when test-takers with dyslexia were provided with extra time (NA Group). These can be observed as differences in slopes in Figure 1.

The t-statistics from the fixed effect regression (using the plm package in R) on the total test, the quantitative part and the verbal part (using equation (1)) gave that $c = 0$ could not be rejected ($p = 0.577$). The same regression gave that $b = 0$ could be rejected ($p = 2.2e-16$) indicating that there is no group difference between NN and AA. Hence, a restricted fixed-effect model (excluding NN) was used to estimate the effect of extended time on each part of the test and the whole test:

$$Y_{it} = a \ast t + b \ast NA_{it} + f_i + e_{it}$$

The effect of extended time, $b$, was estimated at 0.149, 0.173 and 0.125 for the total score, the quantitative score and the verbal score respectively (Table 4). The $a$ coefficient can be interpreted as the general retest effect.

$b > 0$ indicates that extra time helped test-takers with dyslexia to overcome some construct-irrelevant barrier. In coherence with the findings presented by Camara et al. (1998) and Ziomek and Andrews (1998), the improvement of the total score for the NA-group $(a + b)$ was more than three times larger than what could be expected as a general retest effect. The general retest effect is effect is 0.062/0.45 $\approx$ 0.14 standard deviation units (standard deviation of the test was 0.45). That is in line with, but at the lower end of, the retest effects reported by Cliffordson (2004).

![Figure 1. Quantitative, verbal and total score from two successive test administration sessions (autumn 2017 and spring 2018). The score range is from 0.00 to 2.00. NN is the group taking the test without accommodations on both occasions, AA is the group taking the test with accommodations on both occasions and NA is the group that first take the test without accommodations and later with accommodations.](image)

Table 4. Regression analysis summary for estimating the effect of extended time, b, on total score, quantitative score and verbal score using fixed effect model, equation (2).

|                      | Estimate | Std. Error | t-value |
|----------------------|----------|------------|---------|
| **Total score**      |          |            |         |
| General retest effect| 0.062*   | 0.001      | 54.531  |
| Effect of extended time | 0.149*  | 0.013      | 11.052  |
| **Quantitative score** |        |            |         |
| General retest effect| 0.076*   | 0.002      | 47.468  |
| Effect of extended time | 0.173*  | 0.019      | 9.203   |
| **Verbal score**     |          |            |         |
| General retest effect| 0.049*   | 0.002      | 32.540  |
| Effect of extended time | 0.125*  | 0.018      | 7.036   |

*Significant at 0.001 level.
**Group Differences in Speededness due to Extended Time**

Differential speededness was analysed by checking group differences in the relationship between the proposed speededness indicators (omitted items and DIF) and item position.

*Group differences in omitted items.* Each booklet was analysed for omitted items. The 2018a dataset showed an interesting pattern indicating differential speededness and many omitted items at the end of the first booklet. Due to this finding in the 2018a dataset, additional four datasets (2015b, 2017a, 2017b, and 2019a) were analysed to confirm the results.

*Figure 2* shows the relationship between the proportion of omitted items (vertical axis) and item position (horizontal axis). The proportions are aggregated from five different test administrations. The pattern shown in Figure 2 is almost identical for each of the test administrations. From visual inspection, four interesting patterns appear. First, the amount of test-takers skipping late items is much higher in the first booklet in comparison to the other booklets. Second, test-takers with extra time due to dyslexia skip more items than test-takers taking the standard version of the test in general. Third, at the end of the first booklet the amount of omission is lower amongst dyslectic test-takers provided with extra time in comparison to other test-takers. Fourth, the proportion of omitted items is higher at the end of each booklet.

In Table 5, the correlation between item position and proportion of omitted items for test-takers without extra time $r_{\text{std}}$ is compared to the same correlation for the group provided with extra time $r_{\text{ext}}$. The test of differences in correlations between groups is indicated by the $Z$ value (Cohen et al., 2014; Howell, 2009) and corresponding $p$ value indicates the significance level of rejecting $\rho_{\text{std}} = \rho_{\text{ext}}$, where $\rho$ denotes the correlation in the population.

Correlations between item position and the proportion of omission of the item is lower for test-takers provided with extra time due to dyslexia on most booklets in all test administrations (Table 5). This indicates some group differences in speededness assuming that speededness is related to omitted items at the end of the test (each booklet). The comparatively large proportion of omitted items at the end of the first booklet may potentially be explained by a learning effect. Meaning that test-takers might learn to manage time better due to their experience from the first booklet.

*Differential item functioning indicators.* Figure 3 shows the effect sizes, $\ln(\hat{a})$, for the 2018a test administration session. The Maentel-Haenszel differential item function indicators show no obvious pattern of effect size influenced by item positioning (Figure 3). That is, the effect sizes and numbers of non-negligible do not appear to be higher at the end of each subtest or at the end of the booklet. Only two items, Item18 on the first quantitative booklet and Item 22 on the first verbal booklet is showing non-negligible DIF ($|\ln(\hat{a})| > 0.425$), both favouring test-takers with no accommodation. Analysing all five test administration sessions shows no significant correlations nor visually obvious pattern of DIF depending on item position. The results from the used DIF indicators do not support

![Figure 2](image_url). Mean proportion of omission for each item position (in five test administration sessions between 2015 and 2019), comparing test-takers provided with extra time due to dyslexia (Dyslectic) with test-takers taking the test under standard time (Standard).
Table 5. Correlations, $r$, between item position and proportion of omitted items for the compared groups (std and ext) on five different test administrations, including transformed correlation coefficients, $r'$, $z$ scores and corresponding two-tailed $p$ values.

| Test       | $r_{std}$ | $r_{ext}$ | $r'_{std}$ | $r'_{ext}$ | $z$  | $p$  | sig. |
|------------|-----------|-----------|------------|------------|------|------|------|
| Booklet1   |           |           |            |            |      |      |      |
| 15b        | 0.775**   | 0.722**   | 1.033      | 0.912      | 4.053| 0.000| **  |
| 17a        | 0.779**   | 0.749**   | 1.043      | 0.971      | 2.766| 0.006| **  |
| 17b        | 0.774**   | 0.698**   | 1.030      | 0.863      | 5.204| 0.000| **  |
| 18a        | 0.834**   | 0.696**   | 1.201      | 0.860      | 13.014| 0.000| **  |
| 19a        | 0.783**   | 0.681**   | 1.053      | 0.831      | 8.025| 0.000| **  |
| Booklet2   |           |           |            |            |      |      |      |
| 15b        | 0.712**   | 0.738**   | 0.892      | 0.947      | −1.849| 0.064|      |
| 17a        | 0.557**   | 0.448**   | 0.628      | 0.482      | 5.608| 0.000| **  |
| 17b        | 0.730**   | 0.779**   | 0.929      | 1.043      | −3.559| 0.000| **  |
| 18a        | 0.716**   | 0.734**   | 0.899      | 0.937      | −1.446| 0.148|      |
| 19a        | 0.720**   | 0.578**   | 0.908      | 0.659      | 8.968| 0.000| **  |
| Booklet3   |           |           |            |            |      |      |      |
| 15b        | 0.372**   | 0.416**   | 0.390      | 0.443      | −1.746| 0.081|      |
| 17a        | 0.709**   | 0.621**   | 0.885      | 0.727      | 6.078| 0.000| **  |
| 17b        | 0.481**   | −0.050**  | 0.524      | −0.050     | 17.914| 0.000| **  |
| 18a        | 0.590**   | 0.573**   | 0.678      | 0.652      | 0.979| 0.328|      |
| 19a        | 0.373**   | 0.319**   | 0.392      | 0.331      | 2.217| 0.027|      |
| Booklet4   |           |           |            |            |      |      |      |
| 15b        | 0.522**   | 0.561**   | 0.580      | 0.634      | −1.830| 0.067|      |
| 17a        | 0.407**   | 0.532**   | 0.432      | 0.593      | −6.170| 0.000| **  |
| 17b        | 0.614**   | 0.180**   | 0.715      | 0.182      | 16.641| 0.000| **  |
| 18a        | 0.598**   | 0.315**   | 0.690      | 0.326      | 13.864| 0.000| **  |
| 19a        | 0.612**   | 0.672**   | 0.712      | 0.814      | −3.695| 0.000| **  |

* Significant at .05 level (two-tailed).
**Significant at .01 level (two-tailed).

Figure 3. ln(α) for all 40 items in each booklet of test administration 18a (spring 2018). Black vertical bars indicate items that reject the null hypothesis supporting that there is an association between item difficulty and group. Grey vertical areas indicate the last item on each subtest within the booklet.
differential item functioning due to item positioning and do not provide support for group differences in speededness.

Conclusions and Discussion

The results indicate that the SweSAT is speeded but it does not conclusively support differential speededness. Additionally, test-takers with dyslexia improved their results when taking the SweSAT with extra time. Hence, the results from this study do not oppose the supposition that the extra time provided may cause unfair advantage. The potentially unfair advantage indicated by omitted items is most apparent on the first booklet, probably because the test-takers had not yet learned to manage their time effectively.

One increasingly-important aspect of test fairness is addressed in this study, namely the potential conflict between accessibility and test bias. Bias is caused when construct-irrelevant components or construct underrepresentation differentially affect the performance of different groups (American Educational Research Association [AERA] et al., 2014). The presumption when providing extra time for test-takers with dyslexia is that these test-takers are otherwise unable to perform in accordance with their maximum potential. Research has shown that most tests with time limits are speeded (Sireci et al., 2005). Hence, the extra time provided might cause bias to the disfavour of test-takers without extra time. This dilemma is well recognised in previous research (Lovett, 2010). Using the structure proposed by Sheinker et al. (2004), test adaptations can be classified as accommodations that do not alter the test construct, accommodations that might alter the test construct or modifications that do alter the test construct. I am using the following three statements to connect the research questions in this study to the maximum potential rationale and the categorisation of adaptations.

- If the standard version of the test is speeded, then the extra time may be considered to be categorised as a problematic, type-two accommodation or type-three test modification in accordance with Sheinker et al. (2004).
- If test-takers perform better with extra time, then extra time is helping the test-takers to overcome a construct-irrelevant barrier and perform closer to their maximum potential.
- If the test is equally speeded for test-takers with and without extra time, then the extra time can be considered as a type one, or unproblematic, type two accommodation.

The if-condition in the statements are hypotheses that are and can be supported or rejected by empirical evidence. The conclusions in the statements are supported by, and use, definitions from previous theories and frameworks. Agreeing to the proposed statements, extra time can be considered to be an unbiased accommodation in two scenarios: (1) The standard version of the test is not speeded and the test-takers with dyslexia improve their results with extra time. (2) The test is equally speeded for all test-takers and the test-takers with dyslexia improve their results with extra time.

The results of this study indicate that the test is speeded and that test-takers with dyslexia improve their results due to the extra time provided (the improvement is three times larger than what would be expected without the extra time). The results are not conclusive regarding whether the test is differentially speeded when extra time is provided as an adaptation of the test. Scenario one can be ruled out, but since the results are inconclusive on whether extra time cause differential speededness, scenario two is still possible.

It is important to note that the results from this study can be used as evidence for the beneficial effect of extended time for students with dyslexia, but it does not reject the hypothesis that all students would benefit equally from extra time if provided. Nevertheless, substantial DIF would be expected to covary with item position if extra time substantially benefitted all test-takers. However, there is a consistent pattern of group differences in omitted items at the end of each, individually-
timed booklet that may indicate some bias due to extended time. If the effect of speededness on omitted items is larger when the test-taker is unfamiliar with the test format and time restrictions so that greater group differences would be expected on the first booklet – as the analysis also shows.

In addition to the possibility that the inconclusive proof for differential speededness is caused by actual undifferentiated speededness, some limitations might be worth mentioning. First, the correlation between change in item position and item omission may potentially be influenced by other factors (e.g., lack of motivation, exhaustion and test taking strategies). However, motivation should be quite high on average throughout the entire test and the test is designed not to be overly exhausting. Second, care must be taken when defining early and late items as more or less speeded since this continuum has no solid empirical rational. In this study all the late items in each booklet are items within one subtest of the quantitative and verbal part respectively. There is a possibility that these items do not represent the test construct as a whole. This may potentially cause under or over-estimations of the indicators used for analysing differential speededness. A third, general limitation is that the results rest on some strong assumptions. For example, random assignment when estimating the effect of extended time for test-takers with dyslexia, These assumptions may be further investigated. However, all assumptions are supported by evidence and/or experience. Finally, real data from a high-stakes test is used in this study and this delimits the methods that can be used in a validity evaluation process. For example, experimental data and technology-enhanced observations of test-taker behaviour cannot be used and it is not possible to separate the accommodation effect and the disability effect.

The argumentation that can be based on this study is necessary but not sufficient for a complete fairness evaluation of accommodations with extra time. From the SweSAT perspective, the accommodation used would preferably need more supportive evidence on the magnitude of group-independent speededness. It is, for example, plausible that the variation of perceived speededness within the group of test-takers with and without extra time causes overlaps in a way that a great number of test-takers without extra time would need extra time and that some test-takers provided with extra time have more than enough time. One interesting approach would be to combine the benefits from experiments and a real high-stakes testing situation to collect high-stakes experimental data in order to create a double session scenario. I would also suggest research on predictive validity, perceived fairness, response time, group independent test construct and a comprehensive DIF analysis with dyslectic test-takers as the focus group.

Acknowledgements
The author is grateful to Magnus Wikström for his valuable input. The author is also grateful to Per Erik Lyrén, Ewa Rolfsman, Christina Wikström, and anonymous reviewers for valuable comments and suggestions.

Disclosure Statement
No potential conflict of interest was reported by the author(s).

References
American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). Standards for educational and psychological testing. American Educational Research Association.
Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences. Lawrence Erlbaum Associates Publishers.
Camara, W. J., Copeland, T., & Rothschild, B. (1998). Effects of extended time on the SAT® I: Reasoning test score growth for students with learning disabilities. Research Report No. 1998-7. College Board.
Cawthon, S. W., Ho, E., Patel, P. G., Potvin, D. C., & Trundt, K. M. (2009). Multiple constructs and effects of accommodations on accommodated test scores for students with disabilities. Practical Assessment, Research & Evaluation, 14(18), 2. https://doi.org/10.7275/ktkv-3279
Rindler, S. E. (1980). The effects of skipping over more difficult items on time-limited tests: Implications for test validity. *Educational and Psychological Measurement, 40*(4), 989–998.

Schmitt, A. P., Dorans, N. J., Crone, C. R., & Maneckshana, B. T. (1991). Differential speededness and item omit patterns on the SAT. *ETS Research Report Series, 1991*(2), 1–21. https://doi.org/10.1002/j.2333-8504.1991.tb01417.x

Secolsky, C. (1989). Accounting for random responding at the end of the test in assessing speededness on the test of English as a foreign language. *ETS Research Report Series, 1989*(1), 1–26. https://doi.org/10.1002/j.2330-8516.1989.tb00337.x

Sheinker, A., Barton, K. E., & Lewis, D. M. (2004). *Guidelines for inclusive test administration* 2005. CTB/McGraw-Hill. Retrieved March 6, 2019 from http://flraen.org/resources/documents-library/doc_download/787-guidelines-for-inclusive-test-administration

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457–490. https://doi.org/10.3102/00346543075004457

Stafford, R. E. (1971). The speededness quotient: A new descriptive statistic for tests. *Journal of Educational Measurement, 8*(4), 275–277. https://doi.org/10.1111/j.1745-3984.1971.tb00937.x

Stenlund, T., Eklöf, H., & Lyrén, P.-E. (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. https://doi.org/10.1080/0969594X.2016.1142935

Wiberg, M. (2009). Differential item functioning in mastery tests: A comparison of three methods using real data. *International Journal of Testing, 9*(1), 41–59. https://doi.org/10.1080/15305050902733455

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1*(2), 281–295.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Lawrence Erlbaum Associates, Inc.

Ziomek, R. L., & Andrews, K. M. (1998). ACT assessment score gains of special-tested students who tested at least twice.

Zuriff, G. (2000). Extra examination time for students with learning disabilities: An examination of the maximum potential thesis. *Applied Measurement in Education, 13*(1), 99–117. https://doi.org/10.1207/s15324818ame1301_5