<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *International Journal of Testing*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

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

Wiberg, M., von Davier, A A. (2017)

Examining the Impact of Covariates on Anchor Tests to Ascertain Quality Over Time in a College Admissions Test

International Journal of Testing, 17(2): 105-126

<https://doi.org/10.1080/15305058.2016.1277357>

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-137416>

Running head: *Quality Control in a College Admissions Test*

Examining the Impact of Covariates on Anchor Tests to Ascertain Quality Over Time in a College Admissions Test

Marie Wiberg¹
Alina A. von Davier²

¹Professor, PhD, Department of Statistics, USBE, Umeå University, Sweden

² PhD, ACT, Vice President, ACTNext, Research and Development, ACT

Corresponding author

Professor Marie Wiberg
Dept. of Statistics, USBE
Umeå University
SE-901 87 Umeå
Sweden

Acknowledgements

The authors thank Yi-Hsuan Lee for the cross-check of the R-code for the prediction interval, and Jiahe Qian, Deping Li, and Joseph Grochowalski for their suggestions on previous versions of the manuscript. The work presented here reflects the opinions of the authors and not those of ACT. This work was conducted with support from the Swedish Research Council (grant 2014-578).

This paper has been published. Please cite as:

Wiberg, M. & von Davier, A. A. (2017). Examining the impact of covariates on anchor tests to ascertain quality over time in a college admissions test. *International Journal of Testing*. 17, 105-126. <https://doi.org/10.1080/15305058.2016.1277357>

Examining the Impact of Covariates on Anchor Tests to Ascertain Quality Over Time in a College Admissions Test

Abstract

We propose a comprehensive procedure for the implementation of a quality control process of anchor tests for a college admissions test with multiple consecutive administrations. The purpose of this anchor test is to monitor changes in the cohorts of test takers. We propose to examine the anchor tests and each anchor item in connection with covariates to investigate if there was any unusual behavior in the anchor test results over time and if the test results differed for different groups of test takers.

Descriptive statistics, ANOVA, and linear mixed effect models were used to examine the impact of covariates on the anchor test over time. A potential seasonal effect was explored because that could potentially jeopardize the fairness of the test. Descriptive statistics and logistic regression were used to examine the quality of each anchor item over time and to explore the possible impact of covariates. The results are discussed in light of practical recommendations for the methods used and how the results can be used in other large testing programs. It is concluded that examining the impact of the test takers' covariates at different administrations can help in the process of providing fair tests for all test takers.

Keywords: Quality control, linear mixed effect model, logistic regression, item analysis

Introduction

A standardized achievement test can be administered consecutively over several years, thus the quality of the test over time is usually determined from the reliability, validity, and comparability of the test forms. To ensure high reliability, validity, and comparability over time, several aspects of the tests need to be examined, for example, the stability of the difficulty of the test forms, the equating of the test forms, and the steadiness of the composition of the test taker target population and the samples at each administration. The main focus of this study was to implement a quality control process on the set of common items (anchor test) in the Swedish Scholastic Assessment Test (SweSAT) that has been administered on many consecutive administrations without any change in its composition or in its items. In this paper we investigated whether the responses to the common items are invariant across different groups of test takers or if they change over time. Monitoring the quality of the anchor test aids in monitoring the quality of the standardized test forms, which are built to the same specifications as the anchor test. Here we also focused on covariates that are collected at the same time that the tests are administered. These covariates include gender, age and educational background.

The SweSAT shares many similarities with other large testing program; it is given consecutively over time, the items are not known to the test takers in advance, it is a high stakes test as it is used for the admission of the test taker into college. It also has some special features: all items (except the anchor test items) are released the same day as the test is given, the test takers can repeat the test as many times as they wish and only the best test result counts when applying for college, and the same external anchor test is given to a sample of test takers at each administration. The anchor items are administered according to a specific, but variable data collection sampling design, so that the exposure of the items is limited. We believe that the quality control process we propose in this paper can be used for other large-scale testing program if they have an anchor test which is administered at least twice

(consecutively) over time, despite the fact that our example, the SweSAT has very specific features.

Most of the large standardized educational tests today have numerous consecutive administrations over time. Some assessment programs, like the SAT[®] and TOEFL[®] in the US have been administered for the past 40 years. The testing population tends to change over time, and these changes have become more rapid in the recent years due to significant societal events (immigration, access to the Internet, globalization, etc.). These changes might impact how items are answered and how the test is equated. Quality control tools and guidelines (for example, those of the International Test Commission, 2012) are receiving increased interest in the testing community. There is an awareness that we need to consider the test takers' background when performing quality control (Allalouf, 2007). There is, however no direct consensus on how we should keep track of possible changes in the test takers' backgrounds and the connection to specific items and how we should deal with possible changes under which the test scores are still valid. While this study does not provide all the answers, it proposes and illustrates one part of a quality control process that might work for many large testing programs with test forms that are administered more than once a year over many years.

The idea is to emphasize that a good test is a good match of items and people, and therefore, we propose to investigate the samples of test takers, the test, and the items at each administration and across administrations, as it is summarized below, in Figure 1. While some of these analyses are conducted at each administration as part of the operational routine of some testing programs, there are not many examples of a systematic analysis of samples, test, and items over time, according to our knowledge.

[Figure 1 about here]

Literature Review. In this subsection we are reviewing papers that either investigated the quality of the SweSAT or used similar methodologies to those we propose here. Students' background (i.e. covariates) has previously been studied in SweSAT over a two-year time period using a

proportion of test takers who were in a specific group based on age, gender, or education. It was concluded that males and those with higher education had a higher proficiency than the rest of test takers and that older test takers had a higher proficiency on the vocabulary part than the younger test takers (Bränberg, Henriksson, Nyquist, & Wedman, 1990). Also, each year a report is made to examine the composition of the test taker group (for example Stage & Öhman, 2002), and one study focused on examining the test over several years during multiple administrations (Stage & Öhman, 2001). The fact that different test taker groups have different proficiency levels and that the composition of the test taker group might vary over time might influence the equating of the test, especially for the SweSAT because this testing program uses an equivalent group design for equating. The main assumption for SweSAT is that if the items on the anchor test behave similarly over time, then the test probably behaves as intended over time as well. If the anchor test behaves differently over time or if it is affected by different covariates at different time points, it jeopardizes the validity of the equating of the test and other processes that are based on the anchor test. For example, if the demographics of the test takers change significantly over time, the items' properties might be affected. The equating functions may become subgroup dependent due to changes in score distributions across demographic groups or due to the accumulation of errors (see Duong & von Davier, 2012).

In repeated test administrations, it is not uncommon that the composition of the test taker group changes over time. Lyrén and Hambleton (2011) used parts of this SweSAT data material to examine the equivalent group assumption for equating the tests over the years and concluded that this assumption might not always hold due to changes in the test takers' characteristics over time. However, their study focused in particular on examining differences in the mean and standard deviation in the anchor test score over time, but they did not examine if and how item statistics or the anchor test might vary for different test taker groups.

As mentioned above, SweSAT has the exact same anchor test included at each administration, which is more of an exception among large scale testing programs. A similar

concept of an anchor is sometimes used in vertical scaling, where a set of common items are administered across grades (see Kolen & Brennan, 2014, Ch. 9). This particular feature presents a unique opportunity to analyze the anchor test items over time (similarly, the common items in vertical linking are analyzed over grades). The methods presented here can, however, be modified for more traditional tests that have anchor tests that are common and linked across pairs of administrations or tests that have anchors that are linked to an item bank.

In the operational administration of large testing programs, numerous methodologies are employed to investigate the quality of the test results at each administration. These procedures focus on the items, tests, and sample composition. Trend studies are often conducted to ensure that the meaning of the reported test scores is maintained time and to determine if there are any signs of scale drift (for a review see Kolen & Brennan, 2014). In recent years, new types of analyses have been considered for investigating different aspects of quality control of testing programs over time. Some of these methods were adapted from the quality control approaches used in manufacturing, while some others are more familiar in psychometrics but have been used in other types of applications. Our approach here is to integrate several methods for the purpose of investigating different aspects the quality control of a large-scale testing program.

Lee, Liu, and von Davier (2013) suggested the use of linear mixed effect models to detect unusual test administrations. Their work is promising because it seeks to model differences in different administrations, and it provides a useful tool for the test developers. Among other methods, we will use their suggested approach in this paper, although we will examine the possibility of a seasonal effect, which they did not examine.

In the recent years, statistical quality control tools from manufacturing have been implemented to monitor the quality control of the item difficulties or reported test scores for testing programs with an almost continuous administration mode. The idea is to detect trends or abrupt changes in the variable of interest (item difficulty or mean of scaled test scores).

If there are many administrations of the test and if there is an assumption that the test score distribution does not change across time and cohorts (that is, a stationary process) as is the case with college admission tests, one can use quality control charts, such as Shewhart and cumulative sum charts, to determine if the variable of interest is stable over time (a process is in a statistical control state). If there is an accumulation of errors over time or an abrupt change (the process is out of control), for example due to items mistakenly getting revealed in advance for some test takers, it will be visible in these charts. Some of these applications have been discussed in Wang and von Davier (2015) and Lee and von Davier (2013). Time-series techniques, for example, change-point models and hidden Markov models, can also be used for detecting abrupt changes in test scores (Lee & von Davier, 2013). It must, however, be noted that not all of these quality control tools are suitable if there is a correlation over time in the variable of interest (Montgomery, 2009, p. 446). For example, it may be the case that the most knowledgeable and motivated test takers choose to take the test during a particular testing window over the course of a year, while the test takers who take the test for practice may do so during other administrations. This self-selection of the test takers may result in a reoccurring pattern in the observed test scores, that is called, seasonality. Thus, preliminary analyses, such as adjustments for seasonality, need to be conducted on the test score data. See Lee and von Davier (2013) for details.

Perkins, Quaynor, and Engelhard (2011) examined the influence of home language, gender, and social class on mathematical literacy in the Programme for International Student Assessment (PISA) studies. They used the Rasch measurement theory to model test takers' background in order to examine the influence on the items, but they did not examine these aspects over time. This approach could be adjusted for examining the role of covariates over time.

In this paper we focused on the following research questions: Does the composition of the test taker group that takes the anchor test change over the administrations in terms of gender,

education, and age? Is there a seasonal effect in the test administrations? How do the test takers with different backgrounds perform on the anchor test over the time period? Is there a relationship between the achievement level on the SweSAT on the item level and the test taker's background? From previous studies of the SweSAT we hypothesized that males and test takers with higher educational level would do better on the test. We also hypothesized that seasonality could have an impact but no general administration effect should be visible.

Several approaches were used to answer these research questions, including linear mixed modeling, ANOVA, and logistic regression, alongside with descriptive statistics, correlation, Cronbach's alpha, item measures (item difficulty and point biserial correlation).

The rest of this paper is structured as follow. In the next section the college admissions test SweSAT is described, followed by a description of the approaches used to model the test, items, and covariates. The fourth section contains a brief description of the empirical example, and the fifth section contains the results. The paper is concluded with a discussion of practical guidelines and some limitations of the study.

The College Admissions Test

SweSAT is a paper-and-pencil multiple-choice college admissions test with binary responses. The versions used here have 122 items in five sections with correlations with total test score at one administration is given within the parentheses – WORD (Vocabulary, 40 items, .81), READ (Swedish reading comprehension, 20 items, .82), ERC (English reading comprehension, 20 items, .82), DS (Data sufficiency, 22 items, .69), and DTM (Diagrams, tables, and maps, 20 items, .71). The test is given twice a year and is used for selection to higher education. Candidates can take the test as many times as they like, and their highest score is used in the selection process to university. The scores are valid for five years. Over the years, the anchor (common) items have been given to subsamples of the test groups in order to examine the stability of the test scores. Over a time period, $n_{\text{WORD}} = 15$ external anchor items (the correlation of the anchor with total test score is about .74 over all administrations) from the WORD section

(1999 Fall–2004 Spring) were administered, and $n_{DS} = 22$ external anchor items (the correlation of the anchor with total test score is .70 over all administrations) from the DS section (1999 Fall–2008 Spring) were administered over ten years. The data from all test takers who took either the DS anchor test ($N = 36,919$) or the WORD anchor test ($N = 28,950$) were included in this study. Different subsamples of test takers took the two anchors, that is, some test takers took the WORD anchor test and some other test takers took the DS anchor test. The DS anchor test was the same length as the regular DS test, while the WORD anchor test contained substantially fewer items than the regular WORD test. Both anchor tests had acceptable correlations with the total test score as compared with the correlations of each of the subtests with the total test score. Even though the WORD anchor test is very short it correlates quite highly with the total test score (.74 over all administrations and in the interval .69-.76 for each administration). In general, an anchor test should correlate high with the total test score. Although the used anchor tests correlate somewhat lower than recommended in the literature (Dorans et al., 2007) we think that it is still useful to serve as an example for the outlined procedure. The used subtests are used in the testing program and are expected to provide relevant information about the test takers and the test, so it is important for this testing program to study them in detail. Note that these anchor tests were primarily designed to monitor the test and not to be used in the equating process. The equating for this test was conducted using an equivalent groups design where one equates the test in the whole test group and in two reference groups: The first group consists of high schools students and the second reference group is a stratified group with stratification variables being age, gender and education level – the same covariates as used in this study. The equating from the total group is in focus but could be adjusted depending on the equatings in the two reference groups. For more details about the equating process please refer to Lyrén and Hambleton (2011). While this methodology differs from the traditional large scale assessments in the US, it is valid and has been in operational use during 20 years.

Anchor test data and covariates over time

At each administration, the test takers fill out a background questionnaire where the relevant information is their age, gender, and educational background. This information is later used in the equating process and to study different aspects of the test. The possible predictors of interest for this study are summarized in Table 1 followed by a short description and described more extensively below.

[Table 1 about here]

Year is a nominal variable and refers to which year the anchor was administered (coded 1–5 or 1–10) depending on which anchor test was used and *season* is a binary variable that refers to the season of administration (if it was spring then it is coded as 0, and if it was fall then it is coded as 1). *Administration* refers to the 10 or 19 different administrations. Test takers' *gender* is a binary variable coded as 0 for women and 1 for men. Five age levels were used (<21, 21–24, 25–29, 30–39, ≥40 years).

Finished high school education was categorized into eight categories or levels according to the amount of theory in core subjects (mathematics and language) the student had. Level one represents elementary school, level two represents two years of high school, level three represents adult education, level four represents three years of a humanities high school program, level five represents social science and economics high school programs, level six represents natural science and technical high school programs, level seven represents high school programs that aim for a profession, and the last level represents those who did not give information on this variable, i.e. missing data, which was 5.7%. *Ongoing high school* education was categorized into two categories, depending on whether it was ongoing or not. *Finished highest education* and *finished education* were categorized in the same way into six categories (levels), where the lower levels represent categorized high school programs in order of amount of theory in the core subjects of mathematics and language. The first level represents finished elementary school, and the second and third levels represent finished two

years and three years of high school, respectively. The fourth and fifth levels represent having less than two years or more than two years of university studies, respectively. The sixth level contains those who did not provide the information on this variable, i.e. missing data. For *finished highest education* we observed 1.0 % missing data, and for *finished education* we observed 5.4% missing data.

We assumed that the test scores were independent from one test administration to another in this study. Although test takers are allowed to repeat the test, it is very unlikely that there are repeating test takers in this dataset because the anchor test is given in different cities at each administration. We were not aware of any repeaters in the data used here. Repeaters' data needs to be analyzed carefully.

Analyzing the Anchor Test, Items, and Covariates

Test Analysis A good way to examine the quality of a large testing program like SweSAT is to examine the tests and the anchor tests on the total group and on various subgroups with descriptive statistics and to visualize the data. Both correlation and attenuated correlation between the total test scores and the anchor test scores were used to examine the test over time. The reliability of the tests was examined with Cronbach's alpha. There are several guidelines published for Cronbach's alpha; Kline (2000) for example, stated that it should be above 0.7 for operational tests. We followed, however, the guidelines by DeVellis (2003), who stated that below .60 is unacceptable; between .60 and .65 is undesirable; between .65 and .70 minimally acceptable; between .70 and .80, respectable; between .80 and .90 is very good, and much above .90 one should consider shortening the scale.

Modeling the anchor tests and covariates over time

In order to model the anchor test over time with different covariates, both ANOVA and linear mixed effect models can be used because they model the whole anchor test, while controlling for different administrations and different compositions of the test taker group. A linear mixed effect model was included in the analysis because some predictors are naturally

defined as random while others are by definition fixed effects. For an overview of how the predictors were categorized, refer to Table 1.

Let Y_{jkl} be the random variable for the anchor test score of test taker j ($1 \leq j \leq N$) given at administration k ($1 \leq k \leq K$) in season l ($0 \leq l \leq L$), and let $\mathbf{X}_{jkl}'$ be a vector of fixed predictors for test taker j . We can then define a linear mixed effect model as follows:

$$Y_{jkl} = \mu_0 + A_{k(l)} + S_l + \mathbf{X}_{jkl}'\beta + \varepsilon_{jkl}, \quad (1)$$

$$A_{k(l)} \sim N(0, \sigma_a^2), S_l \sim N(0, \sigma_s^2) \quad \varepsilon_{jkl} \sim N(0, \sigma^2),$$

where μ_0 is the intercept or the grand mean of the anchor test score, $A_{k(l)}$ is a random term for administration k in season l . Season S_l is a random term which in general can have many values but in this paper was restricted to two values (0 = spring/1 = fall). β is a vector of coefficients for the fixed predictors $\mathbf{X}_{jkl}'$, and ε_{jkl} is the error term. In the models, administration and season were examined for possible inclusion to examine if they affect the mean proficiency of a test taker. Note the appearance of the linear mixed effect model in Equation 1 can be varied in many ways depending on which effects are significant and by including interactions and other covariates of interest. For quality control in the SweSAT, one should be aware of whether the administration, the year, or the season is significant because this might affect the equating and increase variability over time (and the autocorrelation of the test scores over time). Also, if any demographic variable is a significant predictor this might indicate that some test takers are either advantaged or disadvantaged, which was not intended when constructing the test. Models were constructed by first adding the fixed effects and then subsequently adding random effects. Finally, the interaction terms were added. We aimed to find a model with as high explanatory power as possible while still following the principle of parsimony. In particular, the deviance was used to decide which predictors should be kept in the final models and the intraclass correlation (ICC) was also calculated for the models. There are also other ways to select the best model in linear mixed models and for an overview refer

to Müller, Scealy and Welsh (2013). The assumptions and fit of the chosen final models were examined through residual diagnostics. The residual plots suggested that the assumption of normally distributed residuals hold.

Item Analysis

Each item was examined over time in term of point biserial correlation (r_{pbis}) and item difficulty (p-value) as defined in classical test theory as the proportion of test takers responding correctly to the item. The point biserial correlation should be above .30 and item difficulty should be between .20 and .80 (Kline, 2000). Each item was examined for possible differential item functioning (DIF) over time using the Mantel–Haenszel test as proposed by Holland and Thayer (1988). One can also use logistic regression and random effect models for DIF analysis, but the Mantel–Haenszel test was chosen here due to its simplicity, its widespread use in operational testing, and due to its clear guidelines about DIF items. Two DIF analyses were performed for each anchor item using the anchor total score as the conditioning variable. One DIF analysis compared the two last administrations, and one compared the first and last administrations as it is common in trend studies. As a measure of the effect size of DIF, the Mantel–Haenszel conditional odds ratio $\ln \hat{\alpha}$ was used (Holland & Thayer, 1985). The DIF categorization scheme described by Zieky (1993) for the transformation $-2.35 \ln \hat{\alpha}$, although transformed for $\ln \hat{\alpha}$, was used as in Wiberg (2009):

A_{MH} : Negligible DIF, $\ln \hat{\alpha}$ is not significantly different from zero and $|\ln \hat{\alpha}| < 0.425$.

B_{MH} : Intermediate DIF, either $-0.638 < \ln \hat{\alpha} < -0.425$ or $0.425 < \ln \hat{\alpha} < 0.638$.

C_{MH} : Large DIF, $|\ln \hat{\alpha}| \geq 0.638$.

Modeling items and covariates over time

As previously described, the SweSAT is a multiple-choice test with binary responses. To examine the impact of covariates on the items, we modeled the items within each group (year, seasonality, education, age, and gender) and compared the item parameter estimates. Logistic regression was used to model the items over time, where the dependent variable Z is the

binary (0,1) outcome of each item and $P(Z)$ is the probability of answering an item correctly.

We thus use the following expression

$$\ln[P(Z) / 1 - P(Z)] = b_0 + \mathbf{X}'\beta . \quad (2)$$

Several different models were tried for each item, and the deviance was used as well as the principle of parsimony was followed when choosing the final models. It is also possible to model the items as random effect models and to include season and administration as random effects; in that case, Equation 2 can be extended with a binomial link, but this was not included here. We also acknowledge the possibility of using linear mixed effect models to model the items instead of logistic regression, but logistic regression was chosen here to model the items for illustrative reasons because linear mixed models were used for modeling the whole anchor test. We also acknowledge the possibility to visually inspect the items using item response theory models as the SweSAT contains binary scored items with five options. If such analysis is performed one should use multiple group calibration to account for potentially different subgroups. We excluded this analysis from the paper because it would have required a substantial amount of space.

The SweSAT Example

SweSAT data from several administrations were used. Data manipulations were performed in SPSS 23, and the models described above were fitted to the data with R (R Development Core Team, 2014) using the package lme4 (Bates & Maechler, 2010). Initially, descriptive statistics and an ANOVA were performed with administration as the independent variable and the anchor test score as the dependent variable. Next, different linear mixed models were examined using the possible predictors in Table 1. The selection of predictors in the final model depended on the deviance values of the models and the deviance. The motivation for adding a random seasonal effect was that different types of test takers with different motivations might opt to take the test in the spring versus in the fall. We also examined the possibility of a random effect for administration. All of the other variables were viewed as

fixed independent variables when examining their contributions to the test scores. The chosen final model's assumptions were examined, i.e. to assure that the residuals were normally distributed.

After examining the whole anchor test, the anchor items were examined as described in the method section – i.e. with point biserial correlation, item difficulty, and DIF over administrations – and then with logistic regression.

Results

Results using the whole anchor tests

First, descriptive statistics of the two anchor tests were examined. The reliability scores for the anchor tests were measured with Cronbach's alpha and were in the intervals .65–.76 for WORD and .81–.82 for DS. Comparing these values to the suggested guidelines means that we have either minimally acceptable or respectable Cronbach's alpha values. The correlation between the anchor test and the sum score on the regular test was .70 for DS over all administrations and in the interval 0.65–0.75 for each individual administration. These results are considered good as the practical guidelines indicate that one should have values well above .30. The attenuated correlation for DS over all administrations was 0.80 and in the interval 0.75–0.84 for each individual administration. For WORD, the overall correlation was .74 with an interval of .69–.76 for each individual administration, and the attenuated correlation was .90 with an interval of 0.89–0.94 for each individual administration. Overall, the correlations were quite stable for both WORD and DS over the administrations. Studying the histograms of the score distributions at each administration for each anchor test revealed that the score distributions were approximately symmetrically distributed with mean and standard deviations as seen in Table 2. It is important to note that the means of these anchor tests differ slightly from the means obtained by Lyrén and Hambleton (2011) because they used adjusted mean scores and standard deviations and did not provide the raw mean scores and standard deviations in their tables.

[Table 2 about here]

Figure 2 shows boxplots of the administration time for the two anchor tests, where odd numbers represent fall administrations and even numbers represent spring administrations. In Figure 2 and 3, the boxes represent the interquartile range, i.e. the values between the first and the third quartile. The whiskers represent 1.5 times the interquartile range and the circles represent outliers which are between 1.5 and 3 times the interquartile range away. As it became evident, there is no particular seasonal effect for test takers taking a fall or spring administration (it was also concluded from the linear mixed effect models). There are, however, some fluctuations over the years that should be examined further.

[Figure 2 about here]

Two interesting findings appear when we examine the overall mean of the test across administrations divided into age groups. From Figure 3, it is evident that older test takers performed better on the WORD anchor test and performed worse on the DS anchor test. The first result is in line with the study by Bränberg et al. (1990) and is expected because older people tend to have a larger vocabulary than younger people as learning does not stop at a certain age (Sullivan & Brown, 2014). This trend could also be seen for each test administration. When we examined how male and females performed, it was evident that men performed significantly better on the DS anchor test and slightly better on the WORD anchor test, and these differences were seen across administrations. These plots were not included due to space limitations but they can be obtained from the corresponding author upon request.

[Figure 3 about here]

Examining the mean of the anchor test over the different education categories showed an expected result, i.e. test takers with a higher education level had higher mean scores on the anchor test. For WORD, the highest education level had a mean score of 12.39 and the lowest had a mean score of 8.74, while for DS the highest education level had a mean score of 15.21 and the lowest had a mean score of 9.12. The histograms of the WORD anchor scores over

the different education categories for males and females revealed very similar distributions, and thus the figures are not included in this manuscript. Descriptive statistics and histograms of the DS anchor scores over different education categories for males and females revealed that although the shapes of the distributions were similar, the means were significantly different. The mean DS score of the category with highest education was 16.74 for males but only 13.98 for females. For males within the lowest education category, the mean DS score was 12.02 while for females it was 8.96. These differences were similar across education levels. The histograms are excluded from the paper as they were mainly used to examine the shape of the distributions, but they can be obtained from the corresponding author upon request.

Modeling the anchor test

The F-tests in the ANOVA analyses, with administration time as a factor in the anchor tests for WORD were significant ($F(9, 28940) = 113.00, p = .000$, effect size = 0.035) and for DS ($F(18, 36900) = 22.35, p = .000$, effect size=0.011) were both significant. This is not surprising due to the large sample sizes, and it is important to state that the effect sizes were low. To examine the material more closely, linear mixed models were used on the two anchor tests with different combinations of the predictors. In the final models for both tests, administration was a significant random effect and the fixed effects of maximum education, gender, and age categories were significant for both anchor tests as shown in Table 3. As there was no seasonal effect and no interaction effect between, for example, gender and education, such terms were not included in the final models. For WORD, ongoing high school education was also significant in the model.

[Table 3 about here]

Details about the examined random and fixed effect models for WORD and DS are given in Table 4, especially, the variance of the random effect and its standard errors, the residual variance and their standard error. For both DS and WORD there was a significant difference

of including random effects in the final model compared with not including them. The deviance is smaller, and R^2 are larger when a random effect is used in the final model compared with when only fixed effects are included. Finally, ICC for the final models shown in Table 3 are given in Table 4. Although the contribution from the random effects might appear somewhat low, we believe they should be kept in the final models as the models were improved when they were included in the model as well as for illustration purposes. Examining the residuals in both of the two final models revealed that the residuals were approximately normally distributed.

[Table 4 about here]

If one wants to examine how a future administration works and thus to predict if the process is in control or out of control, a prediction interval for the mean score using the final models in Table 3 can be used, as was done in Lee, Liu, and von Davier (2013). If the estimated mean of a future administration falls outside the prediction interval, the testing procedure should be checked for possible quality control problems. The prediction interval can be defined as

$$\hat{\mu}_{j.} \pm z_{2.5\%} \sqrt{[1 / N] \sigma^2 + \sigma_a^2 + [1 / N^2] [1 \mathbf{X}'_j \text{Var}(\hat{\beta}) \mathbf{X}_j 1]},$$

Where N is the number of test takers in the administrations and σ^2 and σ_a^2 are the variances of the administrations included in the final model and the new administration, respectively. To predict future means, data from the current administrations can be used. For the WORD test, if using the final model in Table 3 a prediction interval for the test score is thus [7.99, 9.95]. If one compares this with the observed means in Table 2, one can note that this interval is somewhat wider than the lowest and highest observed mean values [8.13, 9.66]. In general, prediction intervals are always wide as compared with confidence intervals, but increasing the number of administrations will narrow this interval. If the next observation falls outside of this interval, one should examine the anchor test carefully because this suggests that the anchor test behaves differently than expected.

Quality control of the anchor WORD items

Among the WORD items, only one item had a point biserial correlation smaller than 0.3 at two administrations. Item difficulties varied between 0.2 and 0.91. No items showed any DIF when comparing administrations 10 and 9 and only two items showed intermediate DIF when administrations 10 and 1 were compared. In general, the item characteristics were quite stable over time. In Table 3 the distributions of the covariates are given at two time points for each of the anchor tests. From this one can see that there are some differences in the composition of the test taker groups over time. Logistic regression was used to model the items, and covariates are marked as significant in Table 5 if either the covariate or most of its factor levels were significant. The covariate that affected all the WORD anchor items was maximum education, which was expected. Most of the items were also affected by the test takers' age and gender. Only items 1 and 15 had significant interaction effects between gender and maximum education. Eight items had the variable *Oeduc* as significant, i.e. whether the test taker was part of an ongoing education program or not. Seven WORD items displayed seasonality, and many anchor items were influenced by the administration variable. From a quality-control perspective, it should be noted that the items influenced by seasonality and administration should be examined closely because this might indicate that the items work differently over time and in different seasons.

Quality control of the anchor DS items

Only one DS item had a point biserial correlation lower than 0.3 at two administrations. Item difficulties varied between 0.25 and 0.81, but were in general above .50. Using the Mantel–Haenszel test with previously defined limits, one item had intermediate DIF when administrations 19 and 18 were compared. One item had large DIF and one item had intermediate DIF when comparing administrations 19 and 1. In general, the DS item characteristics were quite stable over time. The items were modeled with logistic regression, and in Table 5 significant factors are marked if either the covariate or most of its factor levels

were significant. The covariates that affected all the DS anchor items were finished highest education, gender, and age. Some items were affected by whether the test taker was in an ongoing education program or not, and a few items (numbers 4, 7, 9, 11, 14, and 22) had significant interaction effects between gender and finished highest education. As seen among the WORD items, a few DS items displayed seasonality and some items were affected according to which administration they belonged to. This observation should be carefully examined because it indicates that the items work differently over time and this could indicate that the test works differently over time. As the composition of test takers changed somewhat between administrations, this could also have influenced the result. For an example of such change in the composition of the covariates refer to Table 3 where the distributions of the covariates are given for two time points for the DS test.

[Table 5 about here]

To better understand the results in Table 5, we examined a random WORD item (number 11) more closely in Table 6. All age categories except age category 2 were significant, and the estimates were higher in the higher age categories, which is in line with what we expect from previous studies of WORD items. Educational levels were, as expected, significant, especially for the highest educational categories four and five. Ongoing education and gender were also significant. For this specific item, female test takers performed in general better than male test takers. There was no year effect on this item, which is good and means that the item worked as expected over the different administration years. From Table 6, it is evident that item 11 has a seasonal effect, and test takers who took the fall administrations performed worse on this item than those who took this item in the spring administrations. This could possibly be explained by the fact that most test takers who take the test in the spring use their test results to apply for university in the fall. Test takers who take the test in the fall have a chance to retake the test in the spring before applying to university the next fall. Despite observing a

season effect on item 11, the seasonality for the whole anchor test has not been found to be significant.

[Table 6 about here]

For quality control issues, it is informative to model each item both across administrations and within each administration. This was done in this study with the 3PL IRT model, but because this requires a significant amount of space the actual plots were excluded here. In general, the DS anchor items were very similar in item discrimination although they differ somewhat in difficulty. The WORD anchor items have much more variation in item discrimination, item difficulty and pseudo guessing. Some of the WORD anchor items appeared to be very easy to guess the correct answer to, as opposed to the DS anchor items that were substantially more difficult to guess and their item characteristic curves were in general quite similar.

Discussion and Concluding Remarks

This study examined different procedures for the examination and monitoring of the quality of the large testing program SweSAT that is given consecutively over multiple administrations and contains external anchor tests. Although our results are specific for the SweSAT, we believe that the approach taken here to examine the quality of the test can be applied to most large testing programs that contain an anchor test which are given at consecutive administrations and when one have access to different covariates. The focus was on examining the stability and quality of the test results and, implicitly, of the equating based on the anchor tests and its items over time. Note we are aware of the special case and use of the anchor test from the SweSAT, but we believe it is useful to conduct the suggested procedure with these common items; they had reasonable values on all item and test measures. Although some of the quality control tools proposed here can be used to examine the equating results explicitly, this was not studied here because Lyrén and Hambleton (2011) already performed such an analysis with the same data material. In this paper, the focus was

instead on discussing the impact of covariates, the time period, and possible seasonality in the anchor items and tests.

The descriptive statistics in Figure 2 and Table 3 showed that age was a major influence on the WORD anchor score, a result that is in line with Bränberg et al. (1990) although they studied a previous version of the SweSAT. It was also evident that educational background had the largest effect in both anchor tests, as shown in Figure 2 and Table 3, and in the specific items, as shown in Table 5. There was no overall seasonal effect on the anchor tests, as shown in Table 3, although some specific items displayed seasonal effects. These seasonal effects mean that these items should be examined closely to assure that their different behavior does not render any unfair advantages for any test takers who repeat the SweSAT. An interesting result was that there was a gender effect in both the WORD anchor test, and in the DS anchor test, as shown in Table 3. Interestingly, the DS test the males had significantly higher means in each of the education categories, and this raises the question of whether high-performing females are actually choosing to take this test. One has to be aware that in order to enter a college in Sweden, you can use *either* this admissions test *or* your high school grades. In Sweden in general, females have much higher grades than males so it is possible that some females have chosen not to take this test because they could enter a university program of their choice without it.

It was clear that a large testing program like SweSAT, which has a number of consecutive administrations, benefits from an investigation of the anchor test and its items together with their covariates to get more information on how the items work over time and with different test taker groups. It was evident from the descriptive statistics and from the modeling of the anchor test and items that there is a relationship between performance and test takers' educational background, age, and gender. As was seen in Table 3, the composition of the test taker group was somewhat different over time.

One limitation with this study was that we could not perform a traditional trend check of equatings over time because the anchor tests were not included in the equating procedure; however, Lyrén and Hambleton (2011) did look at the means and standard deviations of the anchor test over the years and found that there were substantial differences at some of the administrations. If one has an anchor test that is used in the equating procedure of the test of interest, we recommend that one checks whether the equating function from the two first administrations is similar to the equating function between the first and the last administration, as trends in equating functions are identified in traditional studies. One could also investigate the subgroups equating functions for those covariate that seem to have a significant effect on the items and test results.

The main message that can be drawn from this study and that could be applied to other large testing programs are: one should actively perform different quality control methods *both* on anchor tests and the anchor items along with an analysis of the covariates over multiple (consecutive) administrations over several years. This is especially important when the test scores are valid for several years, like the SweSAT scores which are valid for five years. This is also important for other large testing programs, which have anchor tests and collect information about the test takers covariates, such as the Italian Invalsi test, the SAT, and TOEFL. Political decisions or if unemployment suddenly change might affect the test group cohort. The outlined quality control methods can be used to detect unusual behavior in the anchor test results over time and monitor changes in the cohorts of test takers which might influence the anchor test results. It is also important to model the results and the covariates in order to find potential predictors of the test scores. It is especially useful to construct prediction intervals so that we can model how the test score should behave, even if the composition of the test taker group change, as that gives us more reassurance of the predicted behavior of an administration.

References

- Allalouf, A. (2007). An NCME instructional module on quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice*, 26(1), 36-46.
- Bates, D., & Maechler, M. (2010). Lme4: Linear mixed-effects models using S4-classes. R package version 1.1-7.
- Bränberg, K., Henriksson, W., Nyquist, H., & Wedman, I. (1990). The influence of sex, education and age on the scores on the Swedish Scholastic Aptitude Test. *Scandinavian Journal of Educational Research*, 34, 189-203.
- DeVellis, R. F. (2003). *Scale development: Theory and applications*. 2nd edition. Thousand Oaks, CA: Sage Publications.
- N. J. Dorans, M. Pommerich, M. & Holland, P. W. (Eds.) (2007). *Linking and Aligning Scores and Scales*. New York: Springer.
- Duong, M. & von Davier, A. A. (2012). Observed score-equating with a heterogeneous target population. *International Journal of Testing*, 12, 224-251.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale NJ: Erlbaum.
- International Test Commission (2012). *International guidelines on quality control in scoring, test analysis, and reporting of test scores*. [www.intestcom]
- Kline, P. (2000). *The handbook of psychological testing (2nd ed.)*. London: Routledge.
- Kolen, M. J. & Brennan, B. L. (2014). *Test equating, scaling, and linking: Methods and practices*, 3rd edition, New York: Springer.
- Lee, Y-H., Liu, M., & von Davier, A. A. (2013). Detection of unusual administrations using a linear mixed effects model. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods, *New developments in quantitative psychology, Springer proceedings in Mathematics and Statistics*, 66, 133-149.
- Lee, Y-H. & von Davier, A. A. (2013). Monitoring scale scores over time via quality control charts, model-based approaches, and time series techniques. *Psychometrika*, 78(3), 557-575.
- Lyrén, P-E., & Hambleton, R.K. (2011). Consequences of violated equating assumptions under the equivalent group design. *International Journal of Testing*, 36(5), 308-323.
- Montgomery, D. C. (2009). *Introduction to statistical quality control* (6th ed.). Hoboken: Wiley.
- Müller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, 28(2), 135-167.

- Perkins, A., Quaynor, L., & Engelhard, G. (2011). The influences of home language, gender, and social class on mathematics literacy in France, Germany, Hong Kong, and the United States. *IERI Monograph series: Issues and Methodologies in Large-Scale Assessments, Volume 4*, 35-58.
- R Development core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Stage, C. & Öhman, G. (2001). *Högskoleprovets utveckling under åren 1977-2000*. [The test and the test takers group composition and result.] PM report 169. Department of Educational Measurement, Umeå University.
- Stage, C. & Öhman, G. (2002). *Högskoleprovet våren och hösten 2002. Provdeltagarnas sammansättning och resultat*. [SweSAT spring and fall 2002. Test takers group composition and results.] PM report 180. Department of Educational Measurement, Umeå University.
- Sullivan, A. & Brown, M. (2014). *Vocabulary from adolescence to middle-age*. CLS working paper 2014/7. Centre for Longitudinal Studies. Retrieved from www.cls.ioe.ac.uk.
- 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). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Wang, Z. & von Davier, A.A. (2014). *Monitoring of Scoring Using the e-rater Automated Scoring System and Human Raters on a Writing Test*. Research Report 04-14. Princeton, NJ: Educational Testing Service.
- https://www.ets.org/research/policy_research_reports/publications/report/2014/jsek
- 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.

Table 1. Descriptions of the Potential Predictors, with their Labels, Effect type used in the Linear Mixed Models, Levels of the Predictors, and Type of Measurement Scale.

Predictor description	Label	Effect type	Levels	Scale
<i>Time</i>				
Year	Year	Random	5 ^a , 10 ^b	Nominal
Season (spring/fall)	Season	Random	2	Binary
Administration	Adm	Random	10 ^a , 19 ^b	Ordinal
<i>Demographics</i>				
Test taker's gender	Gen	Fixed	2	Binary
Test taker's age	Age	Fixed	5	Ordinal
<i>Educational background</i>				
Finished high school	Fpeduc	Fixed	8	Ordinal
Ongoing high school	Oeduc	Fixed	2	Binary
Finished education	Feduc	Fixed	6	Ordinal
Finished highest education	Meduc	Fixed	6	Ordinal

^aLevels for the WORD test. ^bLevels for the DS test.

Table 2. Means and Standard Deviations of Scores on the WORD Anchor test to the Left and the DS anchor test to the Right.

Adm	Test Season	WORD			DS		
		M	SD	N	M	SD	N
1	1999 Fall	8.51	2.89	1311	12.49	4.77	1843
2	2000 Spring	9.78	2.97	11288	12.29	4.83	2550
3	2000 Fall	9.66	2.93	3919	11.97	4.79	1091
4	2001 Spring	8.35	2.89	1491	12.29	5.06	3499
5	2001 Fall	9.66	3.12	1108	13.11	4.99	3057
6	2002 Spring	8.85	2.98	2112	12.96	4.89	1391
7	2002 Fall	8.13	2.91	1115	11.36	5.02	912
8	2003 Spring	8.89	2.93	4587	12.41	4.92	1080
9	2003 Fall	8.74	2.71	930	12.66	4.93	7216
10	2004 Spring	8.78	2.68	1096	11.75	4.89	1091
11	2004 Fall	-	-	-	12.01	5.14	1543
12	2005 Spring	-	-	-	12.14	5.03	1329
13	2005 Fall	-	-	-	11.52	4.95	1471
14	2006 Spring	-	-	-	11.15	4.89	1190
15	2006 Fall	-	-	-	10.81	5.03	761
16	2007 Spring	-	-	-	12.05	5.09	4156
17	2007 Fall	-	-	-	11.78	4.99	1367
18	2008 Spring	-	-	-	11.70	5.14	624
19	2008 Fall	-	-	-	12.25	4.99	748

Table 3. Fixed Effects in the Final Models for the Anchor Tests WORD to the Left and DS to the Right together with Proportions of Test Takers within each Sample. Random Effects for WORD (Wre) and DS (DSre) in the Lower Part of the Table.

Predictor	WORD			DS						
	Estimate	SE	<i>t</i> -value	99F	04S	Estimate	SE	<i>t</i> -value	04F	08S
Intercept	8.15	0.18	44.41	-	-	10.97	0.22	50.57	-	-
Age2	-0.09	0.05	-1.95	15.1	27.3	-1.17	0.06	-18.89	24.2	23.1
Age3	0.24	0.06	4.11	9.6	9.5	-2.22	0.08	-27.34	10.2	7.4
Age4	1.08	0.06	16.99	12.1	10.6	-2.46	0.09	-27.74	12.4	7.0
Age5	2.30	0.09	25.01	3.9	3.0	-2.91	0.14	-21.08	3.2	2.7
Meduc2	-0.80	0.12	-6.84	12.5	4.0	-0.74	0.19	-3.97	3.5	2.0
Meduc3	0.10	0.11	0.95	77.9	83.4	0.62	0.17	-3.66	82.8	85.6
Meduc4	1.98	0.12	16.86	5.0	6.2	3.27	0.18	18.00	8.5	6.8
Meduc5	2.89	0.13	22.84	2.1	2.8	4.94	0.19	25.22	2.5	3.3
Meduc6	-0.93	0.19	-4.95	0.8	1.2	0.30	0.19	1.01	0.7	0.7
Gender	0.64	0.03	19.53	40.4	48.3	2.62	0.05	53.60	46.2	48.3
Oeduc	-0.10	0.04	-2.48	45.3	35.9	-	-	-	26.3	35.8
Wre	1	2	3	4	5	6	7	8	9	10
	-.21	.72	.59	-.51	.43	-.11	-.58	-.16	-.11	-.07
DSre	1	2	3	4	5	6	7	8	9	10
	.36	.35	.01	.52	1.15	.79	-.11	.26	.53	-.22
	11	12	13	14	15	16	17	18	19	
	-.14	.02	-.71	-.81	-.83	-.25	-.49	-.49	.08	

Meduc = Finished highest education, Oeduc = Indicator variable of whether the test taker was part of an ongoing education program or not. SE = Standard Error, 99F = 99 Fall administration, 04S = 04 Spring administration, 04F = 04 Fall administration, 08S = 08 Spring administration. Wre = The 10 estimates for the random effect for WORD. DSre = The 19 estimates for the random effect for DS.

Table 4. Variances of the Random Effect (varRE), Residual variances (RVar), Standard Errors (SE), Deviances and R^2 for the Final Models with a Random Effect, and the Models when Only Fixed Effects are used together with the ICC for the Final Model.

	varRE Random (SE)	RVar (SE)	Deviance Fixed	Deviance Random	R^2 Fixed	R^2 Random	ICC
WORD	.20 (.45)	7.41 (2.77)	140937	140150	.14	.16	.16
DS	.33 (.57)	20.58 (4.54)	204269	203917	.14	.18	.10

Random = A random effect is included in the model. Fixed = Only fixed effects are included in the model.

Table 5. Significant Variables in the Logistic Regression for WORD Anchor Items to the Left and DS Anchor Items to the Right.

WORD	Season	T*	Meduc	Gen	Age	Oeduc	DS	Season	T*	Meduc	Gen	Age	Oeduc
1		X	X	X			1		X	X	X	X	
2		X	X				2			X	X	X	X
3		X	X	X	X		3	X	X	X	X	X	X
4		X	X		X	X	4			X	X	X	
5	X	X	X	X	X	X	5		X	X	X	X	X
6	X	X	X		X	X	6			X	X	X	X
7		X	X	X	X		7		X	X	X	X	
8			X	X	X	X	8			X	X	X	
9	X		X	X	X		9	X	X	X	X	X	
10	X	X	X	X	X	X	10		X	X	X	X	
11	X		X	X	X	X	11			X	X	X	X
12		X	X	X	X		12			X	X	X	X
13	X		X	X	X	X	13		X	X	X	X	
14		X	X	X	X		14			X	X	X	X
15	X	X	X	X	X	X	15	X	X	X	X	X	
							16			X	X	X	
							17	X		X	X	X	X
							18			X	X	X	
							19			X	X	X	
							20		X	X	X	X	
							21	X		X	X	X	X
							22			X	X	X	

*T refers to either administration or year depending on whether season is significant or not. If Season is significant, it refers to year, otherwise it refers to administration. Gen = Gender, Meduc = Finished highest education, Oeduc = Indicator variable of whether the test taker was part of an ongoing education program or not.

Table 6. Estimates, Standard errors, Z-values, and p -values from modeling WORD item 11 with Logistic Regression.

Variables	Estimate	Std. Error	Z-value	p -value
Intercept	0.307	0.098	3.137	0.002
Age 2	0.045	0.038	1.195	0.232
Age 3	0.447	0.048	9.267	0.000
Age 4	0.752	0.054	13.823	0.000
Age 5	1.252	0.091	13.707	0.000
Meduc 2	-0.167	0.098	-1.697	0.900
Meduc 3	0.002	0.093	-0.020	0.984
Meduc 4	0.636	0.101	6.306	0.000
Meduc 5	0.746	0.116	6.451	0.000
Meduc 6	-1.157	0.148	-7.811	0.000
Oeduc	-0.097	0.032	-2.951	0.003
Gender	-0.264	0.026	10.059	0.000
Season	-0.312	0.029	10.706	0.000

Figure 1. Schematic representation of the proposed QC Process at test and item level with respect to the relevant demographics across many test administrations.

(a) WORD

(b) DS

Figure 2. Boxplots of the test scores for the different administrations (time) of WORD to the left and DS to the right.

Figure 3. Test scores for the different age groups on the WORD anchor test to the left and the DS anchor test to the right.