UiO University of Oslo

Item Performance in Context: Differential Item Functioning Between Pilot and Formal Administration of the Norwegian Language Test

Candidate number: 1111

Master of Science in Assessment, Measurement and Evaluation

30 Credit Points

Faculty of Educational Science, University of Oslo

Center of Educational Measurement in Oslo

May 2020

Table of Contents

Popular Abstract

Different tests are made for different purposes. Some are made for practice, and others can provide results for more serious decision making. Then the motivation and engagement of the test-takers might differ, according to how serious consequences they get from the test results. This difference in test-takers' engagement in different administrations might lead to potential *error* in calculating their test responses. In fact, there are many other sources this *error* might come from, also depending on how the test is made and designed. To detect and analyze this *error* in test responses more systematically, we call this as Differential Item Functioning (DIF). This study explores how the Pilot and Formal tests of Norwegian language test are administered in different ways, and how this can be related to the item performance. Furthermore, we examine how the particular items with influential amount of DIF differ from other items in their features and context.

Acknowledgement

Firstly, my utmost thanks to my extraordinary supervisors, Chia-Wen and Tor! I have felt very fortunate to have you both overseeing my work. Thank you for your patience and engagement all the way, since day one of selecting the topic until the end. Your advice meant a great deal to me, not only in relation to the thesis but also in my learning how to be a professional researcher. You kept challenging me, and I learned so much as a result, which will undoubtedly serve me well in the future.

In looking back on our two years of the master's program, I have many mixed feelings. There were steep learning curves, which involved a great deal of work that was surprisingly tough. In our class, we found ways to help each other with valuable comments, care, and cakes. We made it fun to work together and celebrated each other and the steps in the learning process. Then I finally started to understand things and began to experience the joy of new insights in the educational measurement field. I am enormously grateful to all the excellent team members at CEMO, the amazing teachers, faithful helper Siri, and my dear colleagues for being part of this master's program journey with me. This program made a huge DIF in my life. And here we are, the first graduates of this program in history!

I appreciate all my friends and family for bearing with me while I was immersing myself in the thesis. Thank you, Mom, for all the support you gave, both psychological, emotional, and delicious Korean food. Erlend, thank you for literally suffering with me all along. Thank you for keeping me sane with your patience and care. And above all, thank God, this is done!

Much love, Gaya. Spring 2020

Abstract

The Norwegian language test (No: Norskprøven) administrated by Skills Norway is a high-stakes assessment, the results of which are used by test-takers in various ways. However, the item parameters used in multistage testing in Norskprøven are calibrated from a low-stakes situation, the Pilot test. Potential item parameter shift from the Pilot test to the Formal test might be a concern of practitioners since it reduces the test's reliability and validity. In this study, differential item functioning (DIF) was examined between the Pilot and Formal reading comprehension tests in Norskprøven, using a log-likelihood ratio test method. Purification method was conducted to clean the invariant items and to improve the precision of DIF-item detections. The results revealed 10 DIF items with a large effect size. A different amount of DIF was found in different levels of ability, i.e., non-uniform DIF. DIF items also showed a tendency to vary more in item discrimination than in item difficulty. Lower discrimination parameters in the Pilot test indicated more random error and might be connected to another factor, e.g., low motivation. Regarding the item features and context, i.e., item format and count of words, there was no clear evidence of being related to DIF. However, more items among the anchor items were piloted in two different levels, in contrast to the DIF items. Therefore, test administration and calibration design seem to be more related to the shift in individual item performance rather than to the item features.

Keywords: Item response theory, differential item functioning, item parameter estimation, purification method, multistage test, Norwegian language test, high- and low-stakes assessment, context effect

Item Performance in Context: Differential Item Functioning Between Pilot and Formal Administration of the Norwegian Language Test

Introduction

Test fairness is an important factor in the valid assessment of individuals' true ability and performance shown in the test (American Educational Research Association et al., 2014). McNamara and Ryan (2011) defined test fairness as "*the extent to which the test quality, especially its psychometric quality, ensures procedural equality for individual and subgroups of test-takers and the adequacy of the representation of the construct in test materials and procedures*" (p. 163). It also has a huge practical impact on test validity in the language proficiency test (Zhu & Aryadoust, 2019). Test fairness is threatened when the test results do not reflect the test-taker's ability level, such as in language skills, but are influenced by other factors, for instance, different test situations or cultural backgrounds. This can interfere with the test's validity, and subsequently the fairness of the test (Zhu & Aryadoust, 2019). Therefore, it is important to evaluate the fairness issue in a test that involves continuous monitoring.

One of the ways to evaluate test fairness efficiently is to analyze items if they are functioning differently for different subgroups or test sessions (American Educational Research Association et al., 2014). *Differential Item Functioning (DIF)* occurs when an item has different probabilities that it will be correctly answered conditional on the same ability level for different groups (van der Linden, 2017). The DIF items can lead to biased results between subgroups, such as those with different cultures, different mother tongues, and different educational levels (M. Kim, 2001). Although DIF items do not necessarily indicate a whole study bias, they may contain important information about the subgroups being examined (van der Linden, 2017). In this way, a DIF study is a preliminary step in validating test use by hinting at the potential sources of bias (M. Kim, 2001). Moreover, DIF studies help test developers and users better understand the interaction between test-taker characteristics and test performance.

The Norwegian language test (NO: *Norskprøven*) for adult immigrants is developed and assessed by Skills Norway (www.skillsnorway.no) at the request of the Norwegian Ministry of Education and Research. Norskprøven is a popular large-scale assessment with around 25,000 test-takers each year, and it measures the language proficiency of Norwegian as a second language (Birkeland, Midtbø, & Ulven, 2019). Four different aspects of language proficiency are assessed in Norskprøven: reading, listening, writing, and oral communication.

Reading and listening comprehension tests of Norskprøven are criterion-referenced tests, in which the cut-scores of different levels of the test are carried out with standard-setting procedures (Moe & Verhelst, 2017). To transfer the correct levels of the test from the standard-setting procedure, and to distribute the new items in the correct levels of standards, *Pilot tests* are used for pre-equating. Pre-equating means that the item parameters are estimated beforehand for use in the operational setting (Davey & Lee, 2011). In the Pilot test, new items are tested and item parameters are estimated. Then, these estimates are used to determine which items should be assigned to which levels of difficulties in the *Formal test*.

Pre-equating is more practical than post-equating because the items are already calibrated for the operational use (Davey & Lee, 2011). However, since the items are already located in each position based on the previous calibration, *context effect* is a potential issue (Davey & Lee, 2011). Context effect means that item performance and characteristic are sensitive to the specific way in which it is presented in the test, for instance wording, format, item position in the test, sequencing of the items, and specific features (Davey & Lee, 2011; Leary & Dorans, 1985; Yen, 1980). A question that follows is whether the parameter estimates are tied to the specific context in which an item was pretested or generalized to remain valid across the contexts in which an item may appear in operational tests (Davey & Lee, 2011).

Importantly, the Pilot test of Norskprøven is administered in a low-stakes situation for the examinees. Pilot tests are primarily used as practice by test-takers and are provided at no extra charge with the language courses they are attending. Different engagement in test-taking in high- and low-stakes situations is one of the potential sources of bias in estimation (Pokropek, 2016; Ulitzsch, Davier, & Pohl, 2019; C. Wang & Xu, 2015). In low-stakes situations, examinees experience few or no consequences from their test performance and therefore may not be fully engaged when responding to the test items (Ulitzsch et al., 2019). In such situations, examinees may exhibit the disengaged behavior of omitting responses and randomly guessing. Previous studies suggest that the disengaged behavior of examinees should be regarded as a different construct than proficiency in low-stakes assessments (Pokropek, 2016; Ulitzsch et al., 2019; C. Wang & Xu, 2015). When this is neglected in the measurement procedure, person and item parameter estimates can be biased. Ulitzsch et al. (2019) conceptualized disengaged test-taking behavior by including random guessing and response omission in their hierarchical latent response model. The results pointed out that engagement is related to ability estimates and item parameter estimates.

Moreover, engagement probabilities tend to vary across items, i.e., items evoked disengaged behavior to a different degree. For instance, the probabilities for correct guesses were differently shown on different item formats, such as multiple-choice items and open response items (Ulitzsch et al., 2019). The relation between different item formats and measurement precision is well documented also in several studies. Perkins (1984) assessed several item types and found that the true-false and multiple-choice items produced better test statistics than other types, e.g., missing letters and grammar paraphrase items. A high guessing rate was shown for true-false items, and a low guessing rate for matching, multiple-choice, and constructed responses (Brown & Hudson, 1998).

In the light of previous research findings, our research questions are stated as:

1) How do the items shift from the Pilot test to the Formal test?
    a. Are there influential DIF items with a large effect size?
    b. Nature of DIF items: Direction and behavior of parameter estimates
2) What are the potential factors related to the item shifting from the Pilot test to the Formal test?
    a. Item format, Count of words and Item position

Our hypothesis is that there are influential DIF items between the Pilot and Formal test administration of the reading comprehension test in Norskprøven. We also want to explore potential factors related to the item shifting, by analyzing several features of the DIF items: item format, count of words, and item position or levels. Our hypothesis is that test responses of the items with loaded words might be more affected by disengaged behavior, which can be shown as DIF between the Pilot and Formal tests. We also predict fewer DIF items in multiple-choice format. Additionally, we expect to see any pattern of DIF items in specific position in the Pilot and Formal test.

The remainder of this article is organized as follows. In the next section, the background theories and rationale of the study provide a theoretical framework. Subsequently, the methodology section is presented, including measures, variables, and the analysis procedure. Next, we describe the main results from the analysis by using tables and visualizations. We conclude the paper with a summary of the study, a discussion of the limitations and implications of the results, and suggestions for future research directions.

Theoretical Framework

Background of Norskprøven

The Formal test administration of Norskprøven is a high-stake assessment, which means that the test can be used to make crucial decisions about individuals or aggregated to make decisions about groups (Kingston & Kramer, 2013). The consequences of the test results vary between the test candidates. A certain level of successful certification in Norskprøven is one of the requirements in applying for Norwegian citizenship and admission to higher education in Norway (Skills Norway, 2017; The Norwegian Directorate of Immigration, 2020). Other test candidates, such as Europeans, who do not need the test results for the residence requirement, may still take the test to document their Norwegian language ability, which is a requirement for many jobs.

There are regulations of the duties and rights around taking Norskprøven, which is mandatory for some groups of people (Directorate of Integration and Diversity, 2017). From an internal report from Skills Norway about the results on Norskprøven from 2014 to 2017, some demographic information has been yielded for test-takers (Birkeland et al., 2019). Half the participants had both the entitlement to Norwegian language training free of charge and the obligation to take the test in order to apply for permanent residence or Norwegian citizenship. The rest of the participants had either the entitlement or the obligation (Birkeland et al., 2019). Some test-takers, for example asylum seekers and their family members, were provided free Norwegian language training and Norskprøven in the early stage of their residence in Norway (Directorate of Integration and Diversity, 2017).

The backgrounds of the test candidates for Norskprøven are highly diverse with respect to age, immigration status in Norway, language, and education background (Birkeland et al., 2019; Moe & Verhelst, 2017). For instance, some have a university background, while others have to learn to read and write when they start their Norwegian language courses.

Based on the various backgrounds of the test candidates, it is important that Norskprøven be fair and precise in measuring Norwegian language skills in order to provide the appropriate opportunities and help for those who need it. One of the intentions behind using a multistage test design in the administration of the test is that it would be better suited for a heterogeneous group of test- takers (Moe & Verhelst, 2017).

Test Structure of the Reading Comprehension Test in Norskprøven

IRT methods are applied to many testing objectives for different uses, one example of which is multistage adaptive testing (MST) (Yan, Davier, & Lewis, 2014). Adaptive testing is a test method in which individuals are provided different items that are based on their estimated ability levels on the previous items. MST differs from the traditional adaptive testing in terms of using sets of items, or testlets, rather than single items. The Formal test of Norwegian reading proficiency in Norskprøven is designed as MST. MST has several benefits, such as reduced test length, while maintaining the necessary reliability (Sadeghi & Abolfazli Khonbi, 2017). The routing design of the reading comprehension test in Norskprøven is shown in Figure 1.

The levels of difficulty used for the reading test items of Norskprøven are B2, B1, A2, A1, and under A1, which are built upon the Common European Framework of Reference for Languages (Council of Europe, 2001). A1 is the easiest level, and the difficulty of the test increases to A2, B1, and B2, with B2 being the most advanced level of the test. Different item formats include, for example, short and long multiple-choice items, “click-on-picture”, and “choose the right word, picture or text”. For better understanding, some examples of the practice items are presented in Appendix III (Skills Norway, 2019). However, not all participants are presented with the same items. As described in Figure 1, the test is structured in several stages. The participants are assigned different sets of items based on their results in the previous stage. Sum scores are used to calculate which pretest, main test, and final level each participant suitably falls into. For instance, to start, all participants take Pretest 1. Those who answer correctly on a certain sum of items are guided to the item set of Pretest 2. Those who give a number of correct responses below this sum are provided the Main test A1/A2, and their sum of correct responses for the whole test will determine whether their reading proficiency is at level A2, A1, or Under A1.

However, there are several differences between the Formal and Pilot test structure. Pilot tests are not designed as MST, but they are linear tests with different levels of A1/A2, A2/B2 or B1/B2. The Pilot test does not include the pretest stage. Candidates in the language courses usually choose their own levels or their teachers recommend the appropriate levels.

Item Response Theory

Item response theory (IRT), also known as latent trait theory or item characteristic curve theory, is a measurement perspective that attempts to explain a latent trait among the observed responses within a set of items (de Ayala, 2009). IRT is a statistical modeling process that expresses the relationship between latent characteristics of individuals as predictors and item responses as observed outcome variables. There are several assumptions in unidimensional IRT models (de Ayala, 2009). One is that the probability of correct response is defined as the function of a single latent trait and draws unidimensionality. Therefore, each test is assumed to measure only one latent trait or construct. Another assumption is local independence. An individual’s performance as characterized by the responses on the items are statistically independent of each other, other than their relationship that is attributable to the latent trait(s). Item characteristic curves or item response function is a non-linear logistic function between the probability of correct response and the ability parameter, which is estimated θ (theta) value. Each curve has an intercept and slope that indicate item characteristics. The probability of correct answer suggests different item parameters, depending on which model is chosen as most suitable to the studied response data. In this study, the two-parameter logistic model is selected, because the response data is dichotomously scored, and we are interested in DIF in item difficulty and discrimination.

The two-parameter logistic model (2PM) of Birnbaum is defined as follows: when the one-parameter logistic model is modified in such a way that the discrimination parameter α varies across items (de Ayala, 2009):

$$p(x_j = 1|\theta, \alpha_j, \delta_j) = \frac{e^{\alpha_j(\theta-\delta_j)}}{1+e^{\alpha_j(\theta-\delta_j)}} \tag{1}$$

where θ is the person ability parameter, which is Norwegian reading proficiency in this study, and δ_j and α_j are item j's difficulty and discrimination parameters, respectively. The item difficulty parameter indicates the intercept, and the item discrimination parameter indicates the slope in the item response function.

Differential Item Functioning

DIF can vary in its amount and characteristics, and can be distinguished between *uniform* and *nonuniform DIF* (Mellenbergh, 1982; Swaminathan & Rogers, 1990). Uniform DIF exists when there is no interaction between ability level and group membership. This means that the probability of answering the item correctly is greater for one group than the other uniformly over all ability continuum. Nonuniform DIF exists when there is interaction between ability level and group membership, that is, the difference in the probabilities of a correct answer for the two groups is not the same at all ability levels. Uniform DIF is typically shown by parallel item characteristic curves, and nonuniform DIF has nonparallel item characteristic curves that mostly cross each other (Swaminathan & Rogers, 1990).

IRT-based Likelihood Ratio Test

Many approaches can be implemented to analyze invariance in test items. The Mantel-Haenszel (MH) procedure, the simultaneous item bias test (Shealy & Stout, 1993), and the logistic regression method (Swaminathan & Rogers, 1990) are a few examples of the non-IRT approach for DIF analysis that are widely used. In the current paper, the IRT-based likelihood ratio test (LRT) method is adopted, for several reasons (Thissen, Steinberg & Wainer, 1993). First, both uniform and non-uniform DIF (i.e., DIF on item location and discrimination) can be examined by the LRT (Li & Stout, 1996). The Mantel-Haenszel method does not test for non-uniform DIF, although it has been widely used for DIF analysis in practice (Millsap & Everson, 1993).

Second, the LRT allows the various test lengths among test-takers in MST that yields the missing at random in the response data. In MST, examinees are not administered either the same number or sets of items. Therefore, the sum of the correct responses cannot be used to measure the test-takers' ability levels. Instead, participants skip over certain item sets to get to the next level, which leads to missing data by design (van der Linden, 2017). Missing data by design is recognized as missing-at-random, and can be solved with full information maximum likelihood estimation (Glas & Pimentel, 2008). The Mantel-Haenszel and logistic regression methods suffer from the missing-at-random problem, leading to the serious inflation of Type I error rate (Robitzsch & Rupp, 2009). The LRT can be applied for detecting non-uniform DIF of the data with the missing-at-random (Glas & Pimentel, 2008; Yan et al., 2014).

The LRT detects DIF items by comparing the fit of two models: the compact model as the baseline model and the augmented model as the comparing model (Meade & Wright, 2012). In the compact model, it is assumed that the item parameters are the same across the two different groups. In the augmented model, an item being investigated is assumed to have DIF (M. Kim, 2001). The compact and augmented models are compared to investigate whether the augmented model fits the data better than the compact model. G2 is used as the test statistic for significance of the ratio of the likelihoods from the two models:

$$G_2\ (df) = \frac{-2loglikelihood\ (C)}{-2loglikelihood\ (A)} \tag{2}$$

where C is the compact model and A is the augmented model. G2 is distributed as a χ_2 distribution with a degree of freedom equal to the difference in the number of parameters estimated in the two models. If the augmented model fits the data better, the examined item in that model is differentially functioning (M. Kim, 2001).

Purification of Anchor Items

One way to detect DIF items most effectively is to adopt the strategy of *purification of anchor items* in the analysis. In invariance tests, anchor items are a set of invariant items used to link the metrics of two samples (Kopf, Zeileis, & Strobl, 2015). To clean the anchor items, the purification procedure detects DIF items in a preliminary DIF analysis and removes them from the list of anchor items in the main DIF analysis (Lee & Geisinger, 2016). The purification procedure is beneficial for large-scale analysis due to power improvement, but it is rarely used in the examination of language testing because of its highly technical procedure (Jodoin & Gierl, 2001; Lee & Geisinger, 2016). Although DIF analysis has commonly been done in large-scale language assessments, more efficient and practical forms of detecting strategies are demanded due to the constantly developing and changing test forms and systems. The multistage testing form is one of the newly developed and accepted methods in many language proficiency tests (Moe & Verhelst, 2017).

In this study, the LRT analysis is done in two steps. In the first step, the all-others-as-anchors (AOAA), or all-other method, is implemented (Meade & Wright, 2012; W.-C. Wang, 2004). The AOAA approach begins with a baseline or compact model in which the item parameters are constrained to be equal across the sample groups. In the augmented model, each item is analyzed separately while constraining all the other items in the test, except for the particular item being analyzed. Next, the likelihood ratio is calculated between the baseline model and the augmented model. The items with G_2 as significant are flagged as DIF. In the second step of the procedure, all the non-DIF items are used as anchor items. LRT is used for another DIF analysis, but this time anchor items are constrained as invariant in the augmented model. In the compact model, it is again assumed that all items are invariant. G_2 for the likelihood ratio between the two models is calculated, and, as in the first step, the items with significant G_2 are identified as DIF.

Effect Size Consideration

The existence of DIF items is expected in many language proficiency tests with large groups of test-takers (Ferne & Rupp, 2007). Practically, however, the results of hypothesis testing for DIF detection are not useful in an extremely large sample when the effect size is small (S.-H. Kim, Cohen, Alagoz, & Kim, S., 2007). The crucial aspect is to evaluate the effect size of DIF items in a way that detects the most influential items that can suggest the direction for improvement of test quality. The expected score standardized difference (ESSD) is used to provide information about the magnitude of DIF (Meade, 2010). According to the taxometric framework suggested for using effect size measures, ESSD can be used at the item level and is based on the focal group sample data. ESSD is an expected score version of Cohen’s d and computed as the ratio of the difference between the mean expected score (ES) of the focal group and reference group, and the standard deviation (SD) (Cohen, 1992; Meade, 2010).

$$ESSD_i = \frac{\overline{ES}_{(\gamma F)} - \overline{ES}_{(\gamma R)}}{SD_{ItemPooled}} \tag{3}$$

where ES $_{s(\hat{\theta})i}$ and SD$_{ItemPooled}$ is calculated as below:

$$ES_{s(\hat{\theta})i} = \sum_{k=1}^{m} P_{ik}(\hat{\theta})X_{ik} \tag{4}$$

$$SD_{ItemPooled} = \sqrt{\frac{(N_F - 1)SD_{ES(i|\gamma F)} + (N_F - 1)SD_{ES(i|\gamma R)}}{2 * N_F - 2}} \tag{5}$$

In examining DIF using ESSD, the mean differences between the groups are standardized on a metric for which Cohen's d (1992) recommendations about effect size criteria can be directly applied (Meade, 2010). The rules of thumb criteria suggested for small effect size are (<.02), medium (<.50), and large (<.80) for the absolute value of ESSD (Cohen, 1992).

Signed item difference in sample (SIDS) is the average difference in the expected score of people in the focal and reference groups, with equal theta values (Meade, 2010). Unsigned item difference in sample (UIDS) is similar to SIDS, except that the differences in the expected scores are absolute values before calculating them for the average value (Meade, 2010). In fact, UIDS is the hypothetical amount of DIF when it is assumed always to favor one group over the other in nature (Meade, 2010). Thus, comparing SIDS and UIDS gives an indication of the extent to which differences in ESs vary across different ability levels. In other words, when the absolute values of SIDS and UIDS are similar, the DIF in the item tends to favor one group over the other throughout the whole theta scale. D-Max is the maximum SIDS (either positive or negative) in the sample. D-Max is used to see the maximum extent to which any one test-taker in the focal group is affected by DIF.

Method

Sample and Data

Response data from Formal and Pilot tests, test structures, item contents and characteristics, and pre-calibrated and operationally used item parameter values were provided by Skills Norway. For the Formal test sample, reading test responses from May 2019 were used. For the Pilot test, an anonymous sample was selected for several reasons;

first of all, there are the most number of common items between this Pilot test and the Formal test from May 2019 sample, which gives the most possible items for comparison. Another reason is that when only one specific sample of Pilot test is used, it makes the data more stable rather than using all gathered Pilot data from several years. In this case, the items would have a fairly even number of candidates in administration throughout the test. This is not the case when using all the Pilot data from several rounds over the years, from the initial test in 2013 until 2019. There were, in total, 8,050 participants who took the Formal test in May 2019 for 172 items. The total observation for the Pilot test data, on the other hand, was 4,984 for 301 items. Polytomous items were excluded in the Formal test data set, because the Pilot test only consisted dichotomous items, and our interest was in the common items. Items were all dichotomously scored, 1 for correct and 0 for incorrect responses.

Analysis Procedure

In this study, programming language R version 3.6.1 (R Development Core Team, 2019) was used for the data generation and analysis, along with several packages, e.g., mirt (Chalmers et al., 2019) and TAM (Robitzsch, Kiefer, & Wu, 2019). The samples were loaded into R, and data cleaning was conducted separately for the groups in the Formal and Pilot tests.

Firstly, data cleaning was done separately for both groups. Observations with more than 10 omitted responses (NA) were removed from the data in both groups. Then, descriptive analysis was done separately for the groups, using the psych package (Revelle, 2019). Maximum, minimum, mean and median values were calculated for sample size for items and test length for candidates in each of the Pilot and Formal test. Sample size for items was examined to see how many candidates responded per item. Test length for each candidate was presented, because each candidate was presented for different items and test length. Candidates in the Formal test specifically, was presented for items according to the multistage routing structure of the test. Test length included the items that were not seen by the candidates, as scored as 0. This was excluded as missing data in the main DIF analysis. Following, the ggplot2 package was used for visualization of boxplot comparison for proportion of correct responses for items in the Formal and Pilot test (Wickham, 2016). Boxplots, also called as box-whisker diagrams, are useful tools to display the results visually, with much information as median, quartile range, maximum and minimum values, and potential outliers (Field, A., Miles, & Field, Z., 2012)

Next, the total number of common items between the Formal and Pilot groups were identified to be investigated for DIF analysis. IRT models were generated and inspected for model fit and assumptions criteria for the test response data of the Formal, Pilot and merged data set with common items. Models were generated by using the TAM package (Robitzsch et al., 2019). Expected a posteriori (EAP) estimation and Standardized Root Mean Square Residual (SRMSR) indices are used for checking IRT assumptions and construct validation. EAP is used for ability parameter estimation, based on numerical evaluation of the mean and variance of the posterior distribution (Bock & Mislevy, 1982). EAP is also used for single-level IRT reliability measures (Cho, Shen, & Naveiras, 2019; Monroe & Cai, 2015). EAP estimate > .9 is regarded as a reliable model fit for the given data (Bock & Mislevy, 1982). SRMSR is used as a goodness-of-fit index, with several benefits (Maydeu-Olivares, 2013). SRMSR is an average of standardized residuals, therefore is not affected by the number of items. Additionally, the interpretation of the SRMSR is straightforward and intuitive. SRMSR smaller than 0.05 indicates a negligible amount of misfit (Maydeu-Olivares, 2013).

Figure 2 shows the flow chart of the main data analysis for Norskprøven in this study. Firstly, the DIF analysis was conducted with the purification of matching criterion by the log-likelihood ratio test method (LRT) (Kim, 2001). The mirt package was used for the main DIF analysis procedure (Chalmers et al., 2019). The AOAA approach was used to identify the first anchor items (Lee & Geisinger, 2016). Then, a new model was estimated by restricting the previously detected anchor items from the first step. Here, the parameters of the rest of the DIF items in the data set were freely estimated for both groups. Next, the effect size was calculated to identify items with a large magnitude of DIF. This analysis procedure was done to demonstrate the research question 1) a. Effect sizes of individual items were calculated by extracting expected scores from the model. Index ESSD was used to see if the absolute values of ESSD for some of the items were larger than 0.8, according to Cohen’s d criteria for effect size (Cohen, 1992; Meade, 2010). Other item level indices, such as signed item difference in sample (SIDS), unsigned item difference in sample (UIDS), and D-Max were also taken into consideration.

Once the DIF items with a large effect size were detected, each was visualized individually by item characteristic curves (ICC) and item information function. In addition, the test characteristic curve (TCC) and test information function were generated to evaluate the DIF effect in the test as a whole for the Formal and Pilot test groups. This analysis was done in combination with an examination of the SIDS and UIDS indices and individual item parameter estimation, to investigate the research question 1) b.

An ICC is also known as an item response function and is a nonlinear regression function of the item responses conditional on the ability measured by the test (de Ayala, 2009). Ability or proficiency skills is the θ (theta), i.e., a latent variable underlying the item responses for each individual. The theta in this study represents the reading proficiency skills in Norwegian as a second language. ICCs are not dependent on the latent distribution in the population, i.e., number of individuals located at the same ability level. Although latent distribution influences the parameter estimates obtained and following the ICCs. Two item parameters, item difficulty and discrimination, will be used to describe the ICCs, and the 2PL item response model is selected for the data. An ICC will vary in its intercept with the item difficulty parameter and in its slope with the item discrimination parameter.

The mean and variance differences between the Formal and Pilot test groups were visually presented as two different test characteristic curves. The TCC is also called the expected score function. The expected score function denotes the expected value of the item scores, conditional on the latent variable. The expected score is also referred to as the true score, which is the expected value of the sum score, conditional on the latent value. The conditional expectation is shown below:

$$E(Y|\theta) = E\left(\sum_{j=1}^{J} X_j|\theta\right) = \sum_{j=1}^{J} E\left(X_j|\theta\right) = \sum_{j=1}^{J}\sum_{k=1}^{m_j} x_k P_{jk}(\theta). \tag{6}$$

where $X_j|\theta$ is the item score conditional on the latent variable and $Y|\theta = \sum_{j=1}^{J} X_j|\theta$ is the sum score conditional on the latent variable (de Ayala, 2009).

Item information is shown by the information function and tells us how precisely an item or a test can measure the latent variable in a particular area of the latent continuum. Thus, it provides an amount of certainty about measurement precision and is inversely related to the uncertainty, which is the error associated with ability estimates at the ability level (de Ayala, 2009). In other words, we can identify which item has the most precision in measurement and therefore the most information about the latent variable for which amount. The information is graphically illustrated by item information function and test information function. Item information for dichotomous items in a two-parameter model can be expressed as:

$$\mathrm{I_j(\theta) = \alpha_j^2 p_j(1 - p_j)} \tag{7}$$

The total information for a test has its own function, as described below. This is simply the sum of all the item information functions.

$$I(\theta) = \frac{1}{\sigma_e^2(\theta)} = \sum_{j=1}^{L} I_j(\theta) \tag{8}$$

Finally, item features such as item format, count of words, and item position in the Pilot and Formal test were examined with the descriptive analysis. This analysis was done to investigate the research question 2). Notably, the item position revealed items' difficulty levels, since Norskprøven is a multistage testing. Easy items were typically located at the beginning of the test and difficult items were located at the end of the test. Additionally, item position in the Pilot test informed about whether the item was piloted at one level or two different levels, since not all items were piloted at one level. Moreover, same characteristics of those variables of the anchor items were also examined and compared to the DIF items. The comparison of item features between the anchor and DIF items enabled us to gain insight into the potential interpretations of the occurrence of DIF.

Results

Descriptive Statistics

After the data cleaning procedure, the sample size was reduced from 8,050 to 7,815 in the Formal test and from 4,984 to 4,846 in the Pilot test. Total data set for the Formal test was 7815 observations, i.e., item responses from all the candidates, with 152 variables, i.e., different test items. Total data set for the Pilot test was 4846 observations with 301 variables.

Table 1 shows different descriptive statistics for the Formal and Pilot test response data: sample size for items and test length for candidates in the tests. It is worth noting that the Formal and Pilot tests have different structures. As presented in Figure 1, there are sets of pretests in the Formal test administration, which all test-takers have to take before level-specific tests. This makes the maximum sample size for items in the Formal equal to the total sample size, 7,815. For the Pilot test, the maximum sum of responses is low in comparison, at 639. The Pilot test does not use pretests, but test-takers are referred directly to the level-specific tests they wish to take or are recommended to take. There are three levels in the Pilot test: A1-A2, A2-B1, and B1-B2. Students at the A2 and B1 levels are recommended to take the Pilot test at two levels.

The test length in Table 1 indicates the total number of items each candidate has been presented, either responded or not. This includes the items that a candidate has not yet seen, because of the time limitation. The candidates in the Formal test appear to have longer tests than the candidates in the Pilot test, based on the maximum values, mean, and median in Table 1. This is because additional pretests in the Formal test (see Figure 1).

In Figure 3, two boxplots are compared for the proportion of correct responses per item in the Formal and Pilot tests (Field et al., 2012; Wickham, 2016). The proportion of correct responses for the items shown in the tables can be interpreted as the item difficulty in classic test theory, and also the probability of the correct answer to the items, since the items are scored dichotomously, either 1 for correct response or 0 for wrong response. The left-hand side boxplot representing the Formal test has a higher location of the box and a smaller range between the maximum and minimum value of proportion than the Pilot test. The median is higher for the Formal group, around 0.55 for the Formal and 0.5 for the Pilot. The higher proportion of correct responses implies that the items appear slightly easier for the Formal group than the Pilot group. More variation within the sample group in the Pilot test is also shown with a larger range. Nevertheless, the use of proportion of correct responses as a difficulty estimate is usually inappropriate in MST because the items are responded to by different examinees and have different sample sizes (see Table 1). Also, the different sets of examinees may have different ability levels. The proportion of correct responses only provides an idea about the status of responses to items but cannot be used to do cross-group comparison in MST design.

Model Fit Indices

The IRT 2PL models are fitted for the different data sets using TAM package in R (Robitzsch et al., 2019). In addition to the Formal and Pilot tests separately, a data set with 56 common items was also merged for a 2PL model. The reliability and model data fit index, EAP reliability, and SRMSR are shown for those three data sets in the Table 2.

The EAP reliability for the Formal test data is 0.958, which shows a very reliable data and model fit. The Pilot test has 0.901, indicating a pretty reliable and merged data set. The common 56 items show a value of 0.81, which is an adequate reliability level. A shorter test length will lead to lower reliability, which makes it a reasonable value. A cutoff for well-fitting IRT models is SRMSR ≤ 0.05 (Maydeu-Olivares, 2013). The SRMSR for the Formal test is 0.0524, which is slightly larger than the cutoff rule, but still acceptable. The SRMSR is larger, 0.0632, for the Pilot test, which indicates a less than well-fitting model fit. The SRMSR for the merged data set, with common items between the Formal and Pilot tests, is slightly smaller than the Pilot test, at 0.0613.

DIF Analysis with Purification Procedure

In our DIF analysis, the focal group was defined as the Pilot test group, and the reference group was defined as the Formal test group. In the first step of the DIF analysis, the likelihood ratio test (LRT) was conducted for the merged data set of the two-group observations, with their common 56 dichotomous items. The all-others-as-anchor approach (AOAA) was implemented for every single item with the mirt package in R programming language (Chalmers et al., 2019; Meade & Wright, 2012). The item discrimination and difficulty parameters were examined. A total of 15 items were detected as invariant with p-values larger than .05, which failed to reject the null-hypothesis of the LRT method. The rest of the 41 items in the data set had significant p-values smaller than .05, which rejects the null hypothesis and indicates that the items function differently across the groups. The LRT test was again conducted in the second step of purification. This time, instead of the AOAA approach,15 invariant items from the first step were used as anchor items in the second step. Thus, the parameter estimation of anchor items were restricted as invariant in the Formal and Pilot groups. The results identified two more invariant items, which made 17 anchor items and 39 DIF items in total at the end of the final step. The mean difference of the merged group model was 0.092, and the covariance was 1.358. A table of the significant 39 DIF items with information criteria: AIC and BIC, baseline model index, degrees of freedom, and p-values is shown in Appendix IV.

Effect size

The item-level index ESSD as the effect size measure was evaluated. In total, 10 items with a large effect size were identified for their absolute values of ESSD larger than 0.8 (Meade, 2010). The ESSD, SIDs, UIDs, and D-Max for the 10 items can be seen in Table 3. Among the 10 DIF items with a large effect size, the smallest absolute value of ESSD was found on Item 8, with 0.805. Item 2 showed the largest absolute value of ESSD, as 1.376. These items indicate the smallest and the largest difference between the Formal and Pilot test groups, respectively.

Items 7, 8, and 26 have equal absolute values for SIDS and UIDS, while the rest of the items do not. When the absolute values of SIDS and UIDS are similar, the DIF in the item tends to favor one group over the other throughout the whole theta scale (Meade, 2010). When SIDS is negative, the reference group has generally a higher expected score than the focal group. Thus, SIDS and UIDS values show that for Items 7, 8, and 26, the expected scores are higher in the Formal test than in the Pilot test for the whole theta scale. For the rest of the items, Items 2, 16, 48, 51, 53, 54, and 56, the difference between SIDS and UIDS showed that the expected scores are higher for the Formal test at a certain range of the theta levels than for the Pilot test. However, at the other range of theta levels, the Pilot test had a higher expected score than the Formal test. These conclusions can also be confirmed by observing the item characteristic curves in Figure 4.

The largest D-Max is shown for Item 51, with a positive value of 0.486, indicating that for any member of the Pilot test group sample, the expected score in Item 51 is higher than for any member in the Formal test group, with an amount of 0.486 at most. The smallest D-Max is shown in Item 7, with a negative value of -0.381. This indicates that for any member of the Formal test group sample, the expected score in Item 7 is higher than for any member in the Pilot test group, with an amount of 0.381 at most. The maximum values in Items 2, 7, 8, 16, 26, 48, and 56 favor the Formal group, as shown by the negative values, and the maximum values in Items 51, 53, and 54 favor the Pilot group, as shown by the positive values.

Item Characteristic Curves

Figure 4 shows the item characteristic curves of the 10 DIF items with large effect sizes. The latent continuum, theta on the X-axis, represents reading proficiency as a latent variable. ICCs have a Y-axis that contains the probabilities of correct responses, which are between 0 and 1. Curves that increase more rapidly correspond to items that are more discriminating than others. Curves located along the theta scale indicate item difficulty, with the logic that the probability of a correct response on difficult items increases along with the ability level. The solid line represents the Formal test, and the dotted line represents the Pilot test. Item responses are from the merged data set with 56 common items, dichotomously scored, and fitted for 2PL model.

Overall, most of the ICCs in Figure 4 show non-uniform DIF, which is indicated by the curves of the Formal and Pilot tests varying in both steepness and location, as well as the two curves crossing each other. The nonparallel curves show that the DIF in items varies in amount at each ability level. This indicates that the 10 DIF items have different item difficulty and discrimination parameters between the Formal and Pilot tests. The parameter estimates can also be seen on Table 5. In most of the DIF items, the solid curves are steeper than the dotted lines with various amounts, except for the Item 8. A steeper curve indicates a higher discriminating power between the test-takers of different ability levels. Thus, most of the DIF items have a larger discriminating power in the Formal test than in the Pilot test, except for the Item 8. The dotted curve is steeper than the solid curve for Item 8, which indicates a higher discrimination parameter in the Pilot test than in the Formal test.

Item Information Function

Figure 5 presents the curves for the item information function of the 10 DIF items with large effect size. As in ICC, the X-axis denotes θ, i.e. the ability scale. The Y-axis now shows the information function of the theta, I(θ) for each item. Most of the items, again except for Item 8, which corroborates with the ICCs, have higher information curves for the Formal test than the Pilot test. In the 2PL model, the maximum item information is at the same location as the item difficulty parameter on the θ continuum (de Ayala, 2009). Item difficulty is defined as the level of theta that corresponds to a 50% probability of getting the item correct (de Ayala, 2009). The information curves in Figure 5 show a relatively smaller difference in location parameters between the solid curve for the Formal group and the dotted curve for the Pilot group. A slightly larger difference in location can be seen in Items 7, 8, 26, and 51, than the rest of the items. Rest of the items seem to have similar item difficulty parameters between the Formal and Pilot tests, as indicated by the location of the curves. On the other hand, the discrimination power shown by the height of the curves seems to differ significantly between the solid and dotted curves. The solid curves, representing the Formal test group, are much higher than the dotted curves in most of the items, except for Item 8, which has a slightly higher dotted curve than the solid curve. Item 26 shows a higher solid curve than the dotted curve, but the difference is relatively small compared to the other items. Items seem to have much more discriminating power among respondents at a different ability level in the Formal test group than in the Pilot test group, except for Items 8 and 26. It is noteworthy that the Formal test group curve in Item 51 has remarkably high information. Only this item has a range between 0 and 4 on the y-axis, while for the rest of the items, the range is between 0 and 2.5. Items 53 and 54 also have very high information curves for the Formal test groups compared to the other items.

Test Characteristic Curves and Test Information Function

The mean and variance difference are also visually presented as the Test Characteristic Curve (TCC) on the left-hand side in Figure 6. In the DIF analysis using the multipleGroup function in the mirt package on R (Chalmers et al., 2019), the mean and covariance for the reference group are restricted to 0 and 1. This is to say that the mean and covariance values for the Pilot test group, which are 0.092 and 1.358 respectively, indicate the overall differences in mean and covariance between the Formal and Pilot test groups. This leads to an interpretation of that in this sample and, on average, a randomly selected test-taker in the Pilot test group has 0.092 higher reading proficiency skills in Norwegian as a second language. The mean difference can vary from 0.045 to 0.140 in the 95 percentage of Confidence Interval (CI). The covariance for the Pilot test is shown to be 1.358 when the covariance for the Formal test is restricted to 1. This value varies between 1.236 and 1.480 in the 95 percentage of CI. Conditional on different ability levels, however, this is not a huge difference. The difference is slightly better visualized in discrimination, which can be seen by the curves in TCC crossing each other, as in the ICCs.

The test information function on the right-hand side in Figure 6 directly informs the measurement precision of the test (de Ayala, 2009). Both of the curves in the Formal and

Pilot tests show information for the test-takers at an average proficiency level, a width between -2 and 2 on the θ continuum. The test information function showing the peak of the Formal test group (solid line) is located to the right side of 0 on the ability continuum, while the peak of the Pilot test group (dotted line) seems to be located around 0. This indicates the highest measurement precision at this location, and that it is appropriate to use the Formal test for the purpose of identifying students that perform slightly higher than average ability level, while the Pilot test shows more appropriateness of identifying average-level performing students. Moreover, the test information function shows that, in general, the Formal test group has a higher curve and, therefore, higher measurement precision than the Pilot test. This is also inversely related to the standard error of the ability estimates (de Ayala, 2009). Thus, the lower test information indicates higher standard error in the Pilot test than in the Formal test.

Comparison Between DIF and Anchor Items

Table 4 shows the different characteristics of the 10 DIF items: item format, count of words, difficulty level they belong to in the Pilot test, location in the Formal test, proportion of correct responses in the Formal test (P), item discrimination parameter (α), and item difficulty parameter (δ). The DIF items have different item formats and various amounts of word counts that are inconclusive. The levels and locations also seem to be diverse, showing that the DIF items are emerged from all the levels of difficulties and location. One thing to note is that no DIF items are located at the A2/B1 level in the Formal test. The average value of the proportion of correct responses of each item, shown as P in Table 4, seems to be quite coherent. Most of the items have a larger proportion of correct responses in the Formal test than in the Pilot test, except for Items 26 and 56. The Pilot test group have a slightly higher proportion of correct responses than the Formal test for Item 56, and Item 26 has the same values in both groups.

The item discrimination parameters are reasonably high for both groups, but much larger for the Formal groups in most of the items. Exceptions are Item 8 and 26, which have relatively small differences favoring the Pilot test. Items 2 and 48 have the smallest discrimination parameters for the Pilot test group, 0.66 and 0.57, respectively, while Items 51 and 53 have the largest discrimination parameters for the Formal test group, 4.07 and 3.07. These values show a huge difference from the rest of the items. For the difficulty parameters, the items show quite incoherent values. The easier items, Items 2, 7, 8, 16, 26, and 48, have larger difficulty parameters for the Formal test than the Pilot test, while the difficult items, Items 51, 53, 54, and 56, have larger difficulty parameters for the Pilot test than the Formal test.

The DIF item characteristics are better shown in comparison to the item characteristics of the anchor items. In Table 5, the item characteristics and parameters for 17 anchor items are presented. The item numbers in Table 5 are arbitrary and not related to the item numbers of the DIF items in Table 4. Regarding item format, noticeably, there are many multiple-choice format items in anchor items. There are no open-ended responses with long texts. Concerning word counts, the anchor items seem to have generally fewer words than the DIF items. Although there are many items that share one text with several items among the DIF items, i.e., five or more “which person” items share one text consisting of 543 words. This leads to actual word counts per item that are much fewer than 543 words. More anchor items are from the middle proficiency level, which are A2 and B1, than in the pretest and B2, which represent most of the cases for the DIF items.

The proportion of correct responses (P) is taken from the Formal and Pilot test results. Nine of the anchor items have a higher proportion of correct responses than the Pilot test, and the rest of the eight anchor items, vice versa. The item discrimination parameters (α) and difficulty parameters (δ) are estimated from the new merged model restricting the 17 anchor items, and are therefore the same in both the Formal and Pilot tests. All the anchor items show reasonably high discrimination parameters.

However, the most interesting finding is that there are far more items being piloted in two different levels in the anchor items than in the DIF items. Among the 10 DIF items, there are two items that have been piloted in two different levels, Items 7 and 8, both in levels A1-A2 and A2-B1. The rest of the DIF items are piloted in only one level. On the other hand, there are, in total, 14 items that have been piloted in two different levels among the anchor items, which were Item 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16 and 17. Among those, Item 3, 4, 5, 6, 8, 10, 11, 12, 13, 14 and 15 are piloted at A1-A2 and A2-B1 levels, whilst Item 8, 16 and 17 are piloted at A2-B1 and B1-B2 levels. According to the test administration design, only A2 and B1 items can be piloted at two levels: A2 items at A1-A2 and A2-B2 levels, and B1 items at A2-B1 and B1-B2 levels.

Discussion

Results Summary and Conclusion

Prior studies have documented that stake difference can lead to a potential shift in item parameters conditional on the same ability level (Ulitzsch et al., 2019). When disengaged behavior is ignored in low-stake assessment, this can lead to bias in ability estimation. This can potentially be shown as item shift in different stake situations. However, DIF analysis related to this type of item shift is rarely investigated for tests with a multistage testing design (MST). Furthermore, the explanation and exploration of DIF items are far less discussed than identification procedures. Another issue that was proposed in this study was the pre-equating procedure using the Pilot and Formal test. Test administration and item contents were speculated to lead to a context effect related to the potential item parameter shift. Based on theories and hypotheses, we generated two research questions to investigate.

The current study investigated DIF in a reading comprehension test by using several sequences in the procedure. The response data sets from Formal and Pilot tests were analyzed for DIF using the two-step purification method. The IRT-based likelihood ratio test was used to detect the invariant anchor items and significant DIF items. Effect size was considered to find the most relevant DIF items. Finally, the item contents and contexts were examined and discussed for the DIF items.

Figure 3 showed a slightly higher proportion of correct responses in the Formal test than in the Pilot test. This indicated that some of the same items might appear easier for the groups in the Formal test than in the Pilot. However, the items should not behave differently in measuring estimates, as it is one of the crucial assumptions of IRT models (Leary & Dorans, 1985). Results revealed 17 anchor items and 39 DIF items at the second step of the purification method. Among the DIF items, 10 were found to have an absolute value of ESSD larger than 0.8. This result directly answered the research question 1) a. The ICCs and information curves showed that there was non-uniform DIF across most of the items, which means that items function differently in both difficulty and discrimination when conditioning on the same theta level. Although the difficulty was similar between the Formal and Pilot groups, the discrimination difference was large. Items showed generally higher discrimination power for the Formal group than the Pilot group, as shown through the steeper curves. This was also shown for the information functions. Nine of 10 DIF items with a large effect in the Formal test had remarkably taller peaks of information functions than those conditional on the same theta level in the Pilot test. These results addressed the research question 1) b.

The difference in information between the Formal and Pilot tests might be related to the systematic noise in low-stake situations. Low motivation in Pilot tests is reflected in the systematic noise of disengaged behavior, which leads to increased random error in Pilot tests and decreased information. However, we do not have crucial evidence that motivation is the main issue here. Disengaged behavior is only suggested as a potential factor in causing DIF between the Formal and Pilot tests.

Item characteristics and features were also analyzed to examine and discover the potential factors that might be related to the DIF. The methodology used was descriptive, examining 10 DIF items with large effect size and comparing them to 17 anchor items. Item format, count of words, levels in the Formal test, levels in the Pilot test, proportion of correct responses, and item parameter estimates were compared between DIF and anchor items. The most striking finding was that 8 of 10 DIF items with large effects were only piloted at one level. In comparison, most of the 17 anchor items were piloted at two levels. Only 3 of 17 anchor items were piloted at one level, which were Item 1, 2 and 7. Items that were piloted at two levels, were piloted either at A1-A2 and A2-B1 levels, or at A2-B1 and B1-B2 levels. This leads to that items piloted in those levels are estimated in wider scale of ability levels: A2 items in A1, A2 and B1 levels, and B1 items in A2, B1 and B2 levels. The IRT parameters give us information about how the item functions at all ability levels (de Ayala, 2009). An item piloted on a larger scale of proficiency naturally gives more valid and stable parameter values than items piloted in a narrower ability range. Uncertainty would be greater for the scales that do not have many candidates. One would therefore get the most valid item parameters from the calibration with the most candidates in all the theta levels.

However, other item features seemed to be a minor concern related to DIF. There was no clear evidence in item formats and count of words that were clearly related only to DIF items. Based on these findings, we can conclude that item shift can be affected by the test administration rather than item format and count of words. More specifically, items being calibrated in several different levels in the Pilot test are tested in a broader aspect in the theta scale, and therefore have more precise estimation. This responded to the research question 2).

Limitations and Further Implications

Some limitations of the study are worth noting. One is that the single characteristics of items do not explain the DIF results adequately enough, and additional research may help to identify whether combinations of variables, such as those related to one specific type of item format, will correlate consistently with DIF. Furthermore, a tentative attempt at generalization will require further experimental confirmation. Our study does not focus on the elimination DIF-related factor, but rather explores the quality of DIF exhibiting items. Therefore, a further direction of investigation is suggested for the qualitative contents of DIF items and the possible mitigation of the DIF factor. A closer qualitative follow-up is also suggested for the items that were shown to fall prone to DIF in order to identify item features that contribute to item drift from low-stakes to high-stakes test administration. Another limitation is that we only investigated the reading test of Norskprøven. The DIF examination for listening test between the Formal and Pilot groups is worthy of exploration in future studies to improve the quality of the item bank in Norskprøven. Writing and oral communication test in Norskprøven do not operate the Pilot administration, and therefore are not appropriate for this kind of DIF analysis.

However, several implications of the study can be discussed. Firstly, this study confirms that when obtaining parameters in the item bank, it is necessary to update them based on parameter estimation from the Formal test. Otherwise, several items might exhibit DIF because of the different situation. Secondly, it can be recommended that Norskprøven should eliminate or update the parameters of detected DIF items from this study. Otherwise, ability estimation in future Formal tests may be biased. Finally, as we have found 17 clean anchor items in this study, Skills Norway might consider using these items as a set of sample items for developing new items for the future administration of Norskprøven.

Our results corroborate the findings from previous studies showing that stake differences can lead to potential shifts in item parameters. Also, further investigation of item characteristic and features provide the direction of potential context effect in the test administration. This study sheds crucial light on DIF analysis practice in the multistage testing context and large-scale assessment in language testing, which can lead to more efficient test validation and fair assessment in practice.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Birkeland, P., Midtbø, T., & Ulven, C. H. (2019). RESULTATER PÅ NORSKPRØVEN FOR VOKSNE INNVANDRERE 2014–2017. *Skills Norway*, 25.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Brown, J. D., & Hudson, T. (1998). The Alternatives in Language Assessment. *TESOL Quarterly*, *32*(4), 653. https://doi.org/10.2307/3587999

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C. F., Meade, A., Schneider, L., King, D., Liu, C.-W., & Oguzhan, O. (2019). *mirt: Multidimensional Item Response Theory* (Version 1.31) [Computer software]. https://CRAN.R-project.org/package=mirt

Cho, S.-J., Shen, J., & Naveiras, M. (2019). Multilevel Reliability Measures of Latent Scores Within an Item Response Theory Framework. *Multivariate Behavioral Research*, *54*(6), 856–881. https://doi.org/10.1080/00273171.2019.1596780

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155. https://doi.org/10.1037/0033-2909.112.1.155

Council of Europe. (2001). Common European Framework of Reference for Languages. *Cambridge University Press*.

Davey, T., & Lee, Y.-H. (2011). Potential Impact of Context Effects on the Scoring and Equating of the Multistage Gre® Revised General Test. *ETS Research Report Series*, *2011*(2), i–44. https://doi.org/10.1002/j.2333-8504.2011.tb02262.x de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. The Guilford Press.

Directorate of Integration and Diversity. (2017, January 18). *Timer, prøver og fritak*. IMDi. https://www.imdi.no/norskopplaring/timer-prover-og-fritak/

Ferne, T., & Rupp, A. A. (2007). A Synthesis of 15 Years of Research on DIF in Language Testing: Methodological Advances, Challenges, and Recommendations. *Language Assessment Quarterly*, *4*(2), 113–148. https://doi.org/10.1080/15434300701375923

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Los Angeles: SAGE.

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling Nonignorable Missing Data in Speeded Tests. *Educational and Psychological Measurement*, *68*(6), 907–922. https://doi.org/10.1177/0013164408315262

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I Error and Power Rates Using an Effect Size Measure With the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education*, *14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, *18*(1), 89–114.

Kim, S.-H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF Detection and Effect Size Measures for Polytomously Scored Items. *Journal of Educational Measurement*, *44*(2), 93–116. https://doi.org/10.1111/j.1745-3984.2007.00029.x

Kingston, N. M., & Kramer, L. B. (2013). High-Stakes Test Construction and Test Use. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199934874.013.0010

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor Selection Strategies for DIF Analysis: Review, Assessment, and New Approaches. *Educational and Psychological Measurement*, *75*(1), 22–56. https://doi.org/10.1177/0013164414529792

Leary, L. F., & Dorans, N. J. (1985). Implications for Altering the Context in Which Test Items Appear: A Historical Perspective on an Immediate Concern. *Review of Educational Research*, *55*(3), 387–413. https://doi.org/10.3102/00346543055003387

Lee, H., & Geisinger, K. F. (2016). The Matching Criterion Purification for Differential Item Functioning Analyses in a Large-Scale Assessment. *Educational and Psychological Measurement*, *76*(1), 141–163. https://doi.org/10.1177/0013164415585166

Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647–677. https://doi.org/10.1007/BF02294041

Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement: Interdisciplinary Research & Perspective*, *11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831680

McNamara, T., & Ryan, K. (2011). Fairness Versus Justice in Language Testing: The Place of English Literacy in the Australian Citizenship Test. *Language Assessment Quarterly*, *8*(2), 161–178. https://doi.org/10.1080/15434303.2011.565438

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*(4), 728–743. https://doi.org/10.1037/a0018966

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*(5), 1016–1031. https://doi.org/10.1037/a0027934

Mellenbergh, G. J. (1982). Contingency Table Models for Assessing Item Bias. *Journal of Educational Statistics*, *7*, 105–108.

Millsap, R. E., & Everson, H. T. (1993). Methodology Review: Statistical Approaches for Assessing Measurement Bias. *Applied Psychological Measurement*, *17*(4), 297–334. https://doi.org/10.1177/014662169301700401

Moe, E., & Verhelst, N. (2017). Setting Standards for Multistage Tests of Norwegian for Adult Immigrants. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard Setting in Education* (pp. 181–204). Springer International Publishing. https://doi.org/10.1007/978-3-319-50856-6_11

Monroe, S., & Cai, L. (2015). Examining the Reliability of Student Growth Percentiles Using Multidimensional IRT. *Educational Measurement: Issues and Practice*, *34*(4), 21–30. https://doi.org/10.1111/emip.12092

Perkins, K. (1984). An Analysis of Four Common Item Types used in Testing EFL Reading Comprehension. *RELC Journal*, *15*(2), 29–43. https://doi.org/10.1177/003368828401500203

Pokropek, A. (2016). Grade of Membership Response Time Model for Detecting Guessing Behaviors. *Journal of Educational and Behavioral Statistics*, *41*(3), 300–325. https://doi.org/10.3102/1076998616636618

R Development Core Team. (2019). *R: A language and environment for statistical computing.* http://www.r-project.org

Revelle, W. (2019). psych: Procedures for Psychological, Psychometric, and Personality Research. *Northwestern University, Evanston, Illinois.*, *R package version 1.9.12*. https://CRAN.R-project.org/package=psych

Robitzsch, A., Kiefer, T., & Wu, M. (2019). *TAM: Test Analysis Modules* (Version 3.3-10) [Computer software]. https://CRAN.R-project.org/package=TAM

Robitzsch, A., & Rupp, A. A. (2009). Impact of Missing Data on the Detection of Differential Item Functioning: The Case of Mantel-Haenszel and Logistic Regression Analysis.

Educational and Psychological Measurement, *69*(1), 18–34. https://doi.org/10.1177/0013164408318756

Sadeghi, K., & Abolfazli Khonbi, Z. (2017). An overview of differential item functioning in multistage computer adaptive testing using three-parameter logistic item response theory. *Language Testing in Asia*, *7*(1). https://doi.org/10.1186/s40468-017-0038-z

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194. https://doi.org/10.1007/BF02294572

Skills Norway. (2017, May 23). *Norskprøven ved opptak til høyere utdanning—Kompetanse Norge*. https://www.kompetansenorge.no/prover/norskproven-ved-opptak-til-hoyere-utdanning/

Skills Norway. (2019, June 13). *Leseforståelse nivå A2–B1—Kompetanse Norge*. https://www.kompetansenorge.no/prover/norskprove/ove-til-proven/leseforstaelse-niva-a2-b1/

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*, *27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

The Norwegian Directorate of Immigration. (2020, April 7). *Norwegian oral test for those who apply for Norwegian citizenship*. UDI. https://www.udi.no/en/word-definitions/norwegian-oral-test-for-those-who-apply-for-norwegian-citizenship/

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 67–113). Lawrence Erlbaum Associates, Inc.

Ulitzsch, E., von Davier, M., & Pohl, S. (2019). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level nonresponse. *British Journal of Mathematical and Statistical Psychology*, e12188. https://doi.org/10.1111/bmsp.12188

van der Linden, W. J. (2017). Handbook of Item Response Theory, VOLUME THREE Applications. *Chapman and Hall/CRC*, *3*, 609. https://doi-org.ezproxy.uio.no/10.1201/9781315117430

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. https://doi.org/10.1111/bmsp.12054

Wang, W.-C. (2004). Effects of Anchor Item Methods on the Detection of Differential Item Functioning Within the Family of Rasch Models. *The Journal of Experimental Education*, *72*(3), 221–261. https://doi.org/10.3200/JEXE.72.3.221-261

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York.* https://ggplot2.tidyverse.org/

Yan, D., von Davier, A. A., & Lewis, C. (2014). Computerized Multistage Testing; Theory and Applications. *Taylor & Francis Group, LLC*, 532.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, *17*(4), 297–311. https://doi.org/10.1111/j.1745-3984.1980.tb00833.x

Zhu, X., & Aryadoust, V. (2019). *Examining test fairness across gender in a computerised reading test: A comparison between the Rasch-based DIF technique and MIMIC.* *8*(2), 26.

Table 1

Descriptive Statistics by Formal and Pilot Test

	Sample Size for Items		Test Length for Candidates	
	F	P	F	P
Max.	7,815	639	45	40
Min.	1,046	405	26	27
Mean	1,934.56	504.79	37.63	31.64
Median	1,259	503	41	31

Note: Sample size for items = Number of responses to items, Max. = Maximum value per category, Min. = Minimum value per category, F = Formal test, P = Pilot test, Test Length for Candidates = The total number of items each candidate has been presented.

Table 2

Two-Parameter Logistic Model Fit for the Formal, Pilot, and Merged Data sets

	Formal	Pilot	Merged by Common Items
Number of Items	152	301	56
EAP Reliability	0.958	0.901	0.81
SRMSR	0.0524	0.0632	0.0613

Note: EAP = Expected A Posteriori, SRMSR = Standardized Root Mean Square Root of Squared Residuals.

Table 3

Effect Size for the DIF Items with Large Effect Size

	SIDS	UIDS	D-Max	ESSD
Item 2	-0.249	0.253	-0.340	-1.376
Item 7	-0.232	0.232	-0.381	-0.858
Item 8	-0.202	0.202	-0.305	-0.805
Item 16	-0.139	0.145	-0.204	-0.865
Item 26	-0.192	0.192	-0.251	-0.826
Item 48	0.135	0.162	-0.295	0.863
Item 51	0.272	0.292	0.486	0.939
Item 53	0.199	0.221	0.336	0.855
Item 54	0.185	0.197	0.295	0.825
Item 56	0.155	0.163	-0.259	1.127

Note: SIDS = Signed item difference in sample, UIDS = Unsigned item difference in sample, D-Max = Maximum difference in sample, ESSD = Expected score standardized difference. A large effect size denotes that the absolute value of ESSD > .8 (Meade, 2010).

Table 4

Item Characteristics and Parameters for 10 DIF Items

Item	Format	Count of Words	Level(s) in Pilot test	Location in Formal test	P (Formal/Pilot)	α (Formal/ Pilot)	δ (Formal/Pilot)
2	MC short	49	A1-A2	Pretest 1	0.81/0.55	1.96/0.66	-1.11/-0.33
7	MC short	45	A1-A2, A2-B1	Pretest 2	0.69/0.33	2.13/1.28	-0.25/0.65
8	Click word	53	A1-A2, A2-B1	Pretest 2	0.77/0.45	1.32/1.70	-0.77/0.04
16	Choose Text	-	A1-A2	A1/A2	0.69/0.66	2.02/0.85	-1.35/-1.23
26	Calender	26	A1-A2	A1/A2	0.44/0.44	1.45/1.24	-0.62/0.13
48	Voice of opinion	296	B1-B2	B1/B2	0.46/0.37	1.97/0.57	1.29/1.63
51	Which Person?	543	B1-B2	B1/B2	0.73/0.66	4.07/1.21	0.69/-0.13
53	Which Person?	543	B1-B2	B1/B2	0.57/0.54	3.07/0.97	0.98/0.53
54	Which person?	543	B1-B2	B1/B2	0.56/0.50	2.53/0.98	0.99/0.56
56	Which Person?	543	B1-B2	B1/B2	0.34/0.36	2.02/0.63	1.53/1.68

Note: P = Proportion of correct responses, α = Item discrimination parameter, δ = Item difficulty parameter. MC short = Multiple choice in short-form.

Table 5

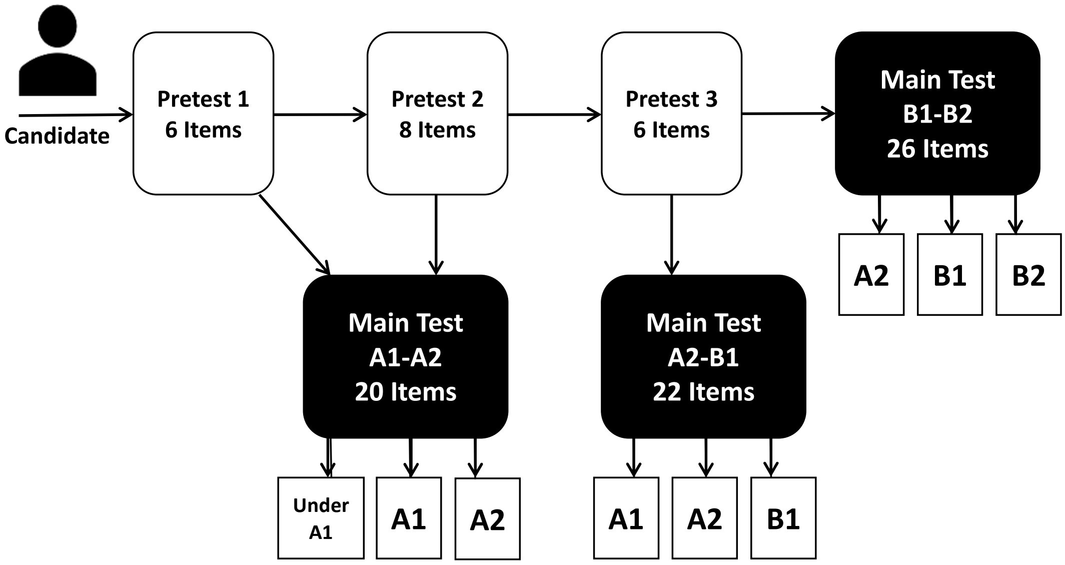
Item Characteristics and Parameters for 17 Anchor Items

Anchor Item nr.	Format	Count of Words	Level(s) in Piloting	Level in Formal Test	P (Formal/ Pilot)	α	δ
1	Click Image	-	A1-A2	Pretest 1	0.87/0.74	1.37	-1.75
2	MC short	31	A1-A2	A1-A2	0.71/0.76	1.66	-1.45
3	Click word	37	A1-A2, A2-B1	Pretest 2	0.74/0.41	1.84	-0.47
4	Click word	49	A1-A2, A2-B1	A2-B1	0.46/0.29	2.29	0.36
5	Click word	49	A1-A2, A2-B1	A2-B1	0.61/0.39	1.70	0.01
6	MC short	62	A1-A2, A2-B1	Pretest 2	0.83/0.58	2.49	-0.70
7	Click word	62	B1-B2	B1-B2	0.55/0.47	0.66	0.84
8	Click word	66	A2-B1, B1-B2	A2-B1, B1-B2	0.44/0.55	1.58	0.30
9	MC short	80	A1-A2, A2-B1	A1-A2	0.32/0.41	0.82	0.19
10	MC short	81	A1-A2, A2-B1	Pretest 2	0.71/0.48	1.20	-0.35
11	MC short	143	A1-A2, A2-B1	A1-A2	0.26/0.35	0.80	0.60
12	MC long	174	A1-A2, A2-B1	A1-A2	0.33/0.57	1.90	-0.39
13	MC long	174	A1-A2, A2-B1	A1-A2	0.45/0.63	1.82	-0.71
14	MC long	174	A1-A2, A2-B1	A1-A2	0.42/0.56	1.32	-0.53
15	MC long	174	A1-A2, A2-B1	A1-A2	0.58/0.67	1.44	-1.06
16	MC long	183	A2-B1, B1-B2	Pretest 3	0.59/0.41	2.03	0.34
17	MC long	183	A2-B1, B1-B2	Pretest 3	0.39/0.28	1.56	0.96

Note: P = Proportion of correct responses, α = Item discrimination parameter, δ = Item difficulty parameter, MC short = Multiple choice in short-form, MC long = Multiple choice in long-form.

Figure 1

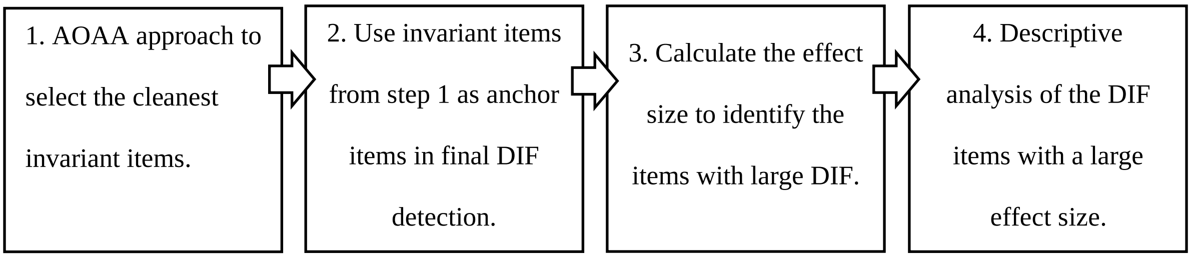
Routing Structure of the Reading Comprehension Formal Test in Norskprøven



Note. Multistage test form of routing structure for Formal administration of reading comprehension test in Norskprøven is provided from Skills Norway. All candidates start the test with the Pretest 1 and the sum of the correct response guides them to next set of items.

Figure 2

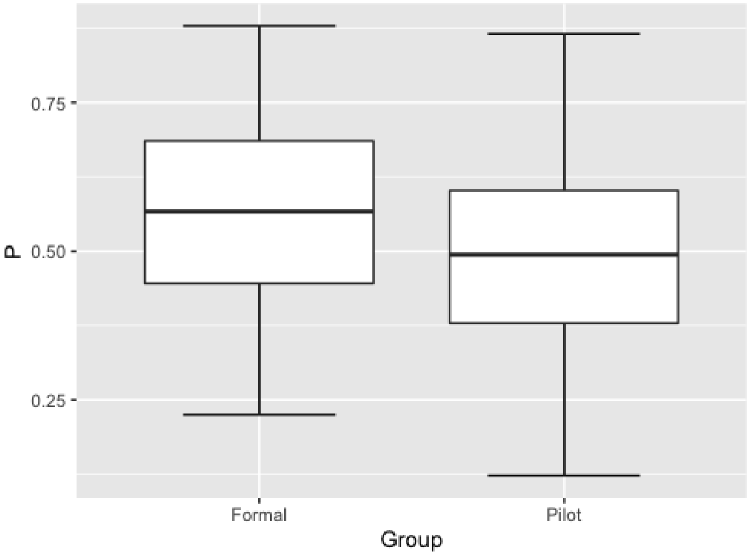
Methodology Procedure


1. AOAA approach to select the cleanest invariant items.
2. Use invariant items from step 1 as anchor items in final DIF detection.
3. Calculate the effect size to identify the items with large DIF.
4. Descriptive analysis of the DIF items with a large effect size.

Note. AOAA approach = All-others-as-anchors approach, DIF = Differential Item Functioning.

Figure 3

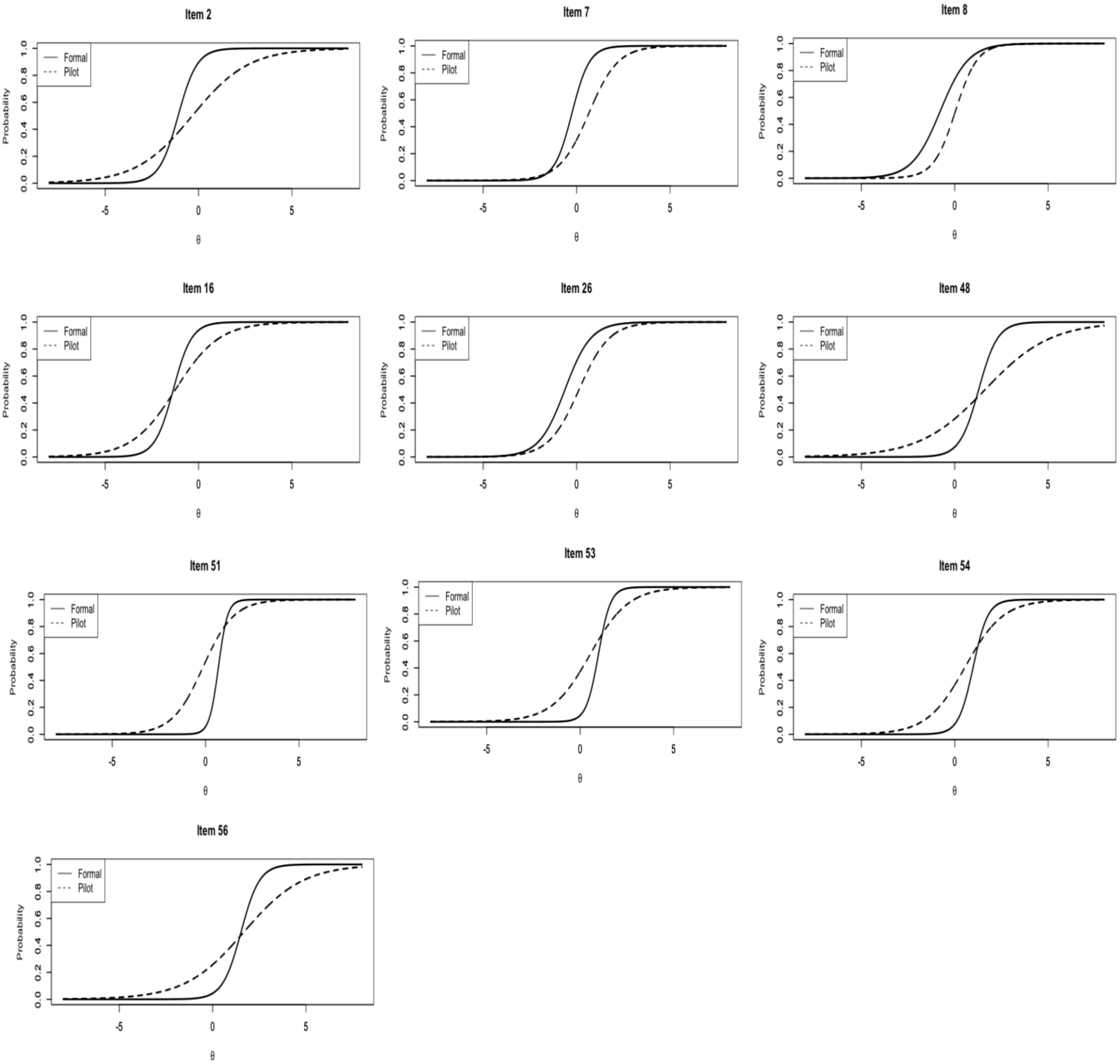
Boxplot Comparison for Proportion of Correct Responses for Items in Formal and Pilot Test



Note. Boxplots for Formal and Pilot test responses in reading comprehension Norskprøven are compared. P = Proportion of correct response.

Figure 4

Item Characteristic Curves (ICCs) by the Two Groups for the 10 Items with Large Effect Size.



Item 2

Item 7

Item 8

Item 16

Item 26

Item 48

Item 51

Item 53

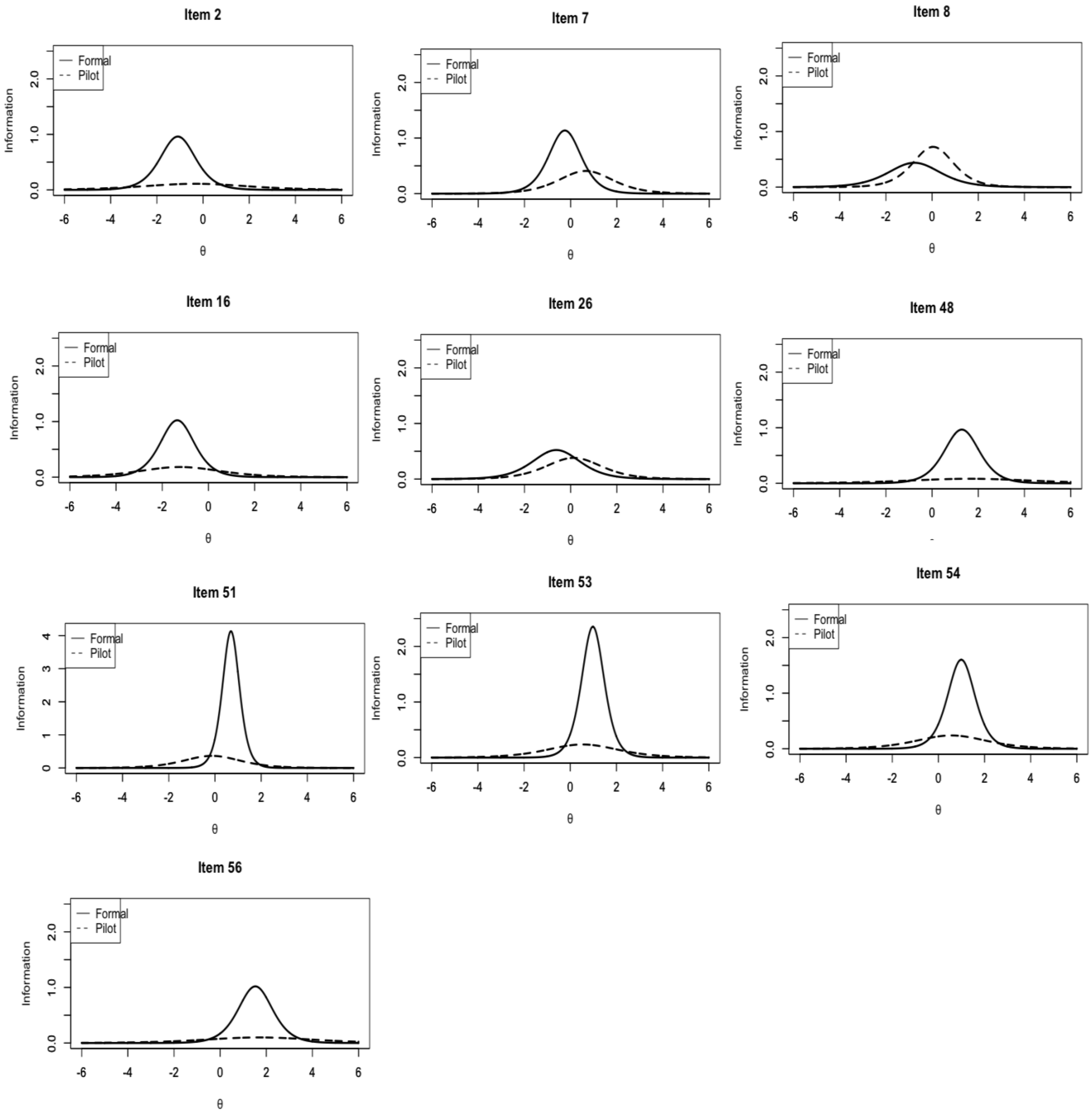
Item 54

Item 56

Note. X-axis = θ (theta), Y-axis = Probability of correct response, Solid line = ICC for Formal test item response, Dotted line = ICC for Pilot test item responses

Figure 5

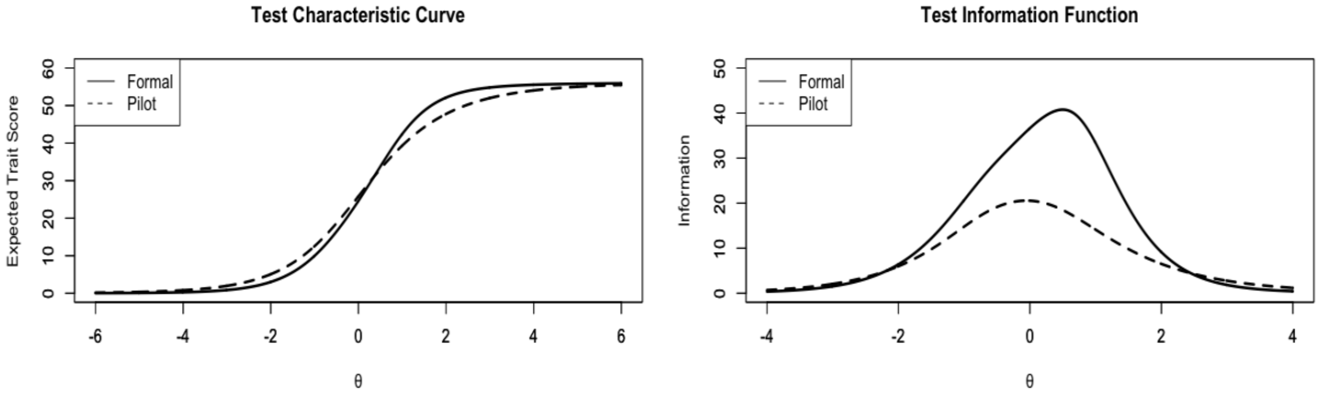
Item Information Curves by the Two Groups for the 10 Items with Large Effect Size.



Note. X-axis = θ (theta), Y-axis = Item Information function, Solid line = Information function curve for the Formal test item, Dotted line = Information function curve for the Pilot test item

Figure 6

Test Characteristic Curve and Test Information Function by the Two Groups.



Note. Test characteristic curves for the Formal and Pilot test on the left side: X-axis = θ (theta), Y-axis = Expected trait score, Solid line = Test characteristic curve for the Formal test, Dotted line = Test characteristic curve for the Pilot test.

Test information function for the Formal and Pilot test on the right side: X-axis = θ (theta), Y-axis = Test information function, Solid line = Information function curve for the Formal test, Dotted line = Information function curve for the Pilot test.

Appendix

Appendix I: GDPR documents & Ethical approval

Application form for Norwegian Centre for Research Data (NSD)

Below is a copy of the application form for NSD filled out on the website. Answers about the thesis project are given in *italic font* under each question.

NOTIFICATION FORM (ENGLISH TRANSLATION) – NSD

NB! First draft

- Personal data
- Types of data
- Project Information
- Responsibility
- Sample and Criteria
- Third Persons
- Documentation
- Other approvals
- Processing
- Information Security
- Duration of project
- Additional Information
- Send in

Which personal data will be processed?

Name
No

National ID number or other personal identification number
No

Date of birth
No

Address or telephone number
No

Email address, IP address or other online identifier
No

Photographs or video recordings of persons
No

Audio recordings of persons
No

GPS data or other geolocation data
No

Demographic data that can identify a natural person
No

Genetic data
No

Biometric data
No

Other data that can identify a natural person
If you think that you will be processing personal data but cannot find a suitable alternative above, indicate this here.
No

Will special categories of personal data or personal data relating to criminal convictions and offences be processed?

Racial or ethnic origin
No

Political opinions
No

Religious beliefs
No

Philosophical beliefs
No

Trade Union Membership
No

Health data
No

Sex life or sexual orientation
No

Criminal convictions and offences
No

Project Information

Edit project Register new project Chose existing project
under 'Register new project':

Title
"Item Performance in Context: Differential Item Functioning Between Pilot and Formal Administration of the Norwegian Language Test"

Project description

The project is a master's thesis about evaluating quality of Norwegian language test. The test response data will be analyzed if there is any bias has been emerged during test calibration from pilot study to formal test.

Subject area

- *Social sciences*

Will the collected personal data be used for other purposes, in addition to the purpose of this project?
No

Explain why it is necessary to process personal data.
-

Project description
Chose file...

External funding
No external funding.

Type of project

- *Student project, Master's thesis*

Responsibility for data processing

Data controller
Skills Norway.

Project leader (research assistant/ supervisor or research fellow/phD candidate)

Name
Position
Email address
Telephone number

Internal supervisor: Chia-Wen Chen.
Postdoctoral Research Fellow at Center for Educational Measurement in Oslo (CEMO)
c.v.chen@cemo.uio.no
mobile: +4747726778

External supervisor: Tor Midtbø.
Advisor at Skills Norway
Tor.midtbo@kompetansenorge.no
mobile: +4747259178

Will the responsibility for processing personal data be shared with other institutions (joint data controllers)?
No

Joint data controllers
-

Whose personal data will be processed?

Sample 1

Describe the sample

Recruitment or selection of the sample

Main analysis will be comparison between Norwegian reading proficiency test responses of formal test and pilot test. Formal test responses are from summer 2018, and pilot test responses are from one particular anonymous period (because of security reason). Test responses from both tests are given by equivalent population; adult immigrants in Norway, but tests differ in their stake-situation. Formal test provides official certificate in Norwegian language, while pilot test has its purpose in practice for the test takers. Pilot test results are also used to calibrate item parameters for test developers.

Age

Adults (18 +)

Will you include adults (18 år +) who do not have the capacity to consent?

No

Types of personal data - sample 1

Name
National ID number or other personal identification number
Date of birth
Address or telephone number
Email address, IP address or other online identifier
Photographs or video recordings of persons
Audio recordings of persons
GPS data or other geolocation data
Demographic data that can identify a natural person
Genetic data
Biometric data
Other data that can identify a natural person

Methods /data sources - sample 1

Select and/or describe the method(s) for collecting personal data and/or the source(s) of data

Data will be provided from Skills Norway, who administers and develops Norwegian language test. Data is test responses for each and every item separately from two different administrations, called; a formal test and a pilot test. Data is already collected, organized and partly cleaned. Only the test responses as pattern of numbers either 0 (wrong) or 1 (correct), are to be used without any personal data nor identifiable keys.

Information - sample 1

Will you inform the sample about processing their personal data?

No – no personal data is prosessed.

How?

Written information (on paper or electronically)
Oral information

Information should be given in writing or electronically. Only in special cases is it applicable to give oral information, if a participant asks for this. See what you must give information about.
Upload information letter
Upload copy of oral information
No

Explain why the sample will not be informed about the processing of their personal data.
+ Add sample

Third persons

No personal data about third persons is processed.

Describe the third persons
Types of personal data about third persons
Name
National ID number or other personal identification number
Date of birth
Address or telephone number
Email address, IP address or other online identifier
Photographs or video recordings of persons
Audio recordings of persons

GPS data or other geolocation data
Demographic data that can identify a natural person
Genetic data
Biometric data
Other data that can identify a natural person

Which sample will provide information about third persons?
Sample 1
Sample 2 etc.

Will third persons consent to the processing of their personal data?
No

Will third persons receive information about the processing of their personal data?
No

Upload information letter
Chose file...
No

Explain why third persons will not be informed.
No personal data is processed.

Documentation

Total number of data subjects in the project
(Data subjects: persons whose personal data you will be processing)
No personal data is used.

How can data subjects get access to their personal data or how they can have their personal data corrected or deleted?
-

Other approvals

Will you obtain any of the following approvals or permits for the project?
No.

- Ethical approval from The Regional Committees for Medical and Health Research Ethics (REC)
- Confidentiality permit (exemption from the duty of confidentiality) from the Regional Committees for Medical and Health Research Ethics (REC)
- Approval from own management for internal quality-assurance and evaluation of health services (intern kvalitetssikring) (The Health Personnel Act § 26)
- Confidentiality permit (exemption from the duty of confidentiality) from the Norwegian Directorate of Health, for quality-assurance and evaluation of health services (kvalitetssikring) (The Health Personnel Act § 29b)
- Biobank
- Confidentiality permit (exemption from the duty of confidentiality) from Statistics Norway (SSB) Statistics Norway has the authority to grant a confidentiality permit for the data that they manage, e.g. data about population, education, employment and social security.
- Approval from The Norwegian Medicines Agency (Statens legemiddelverk, SLV) E.g. for a clinical drugs trial

- Confidentiality permit (exemption from the duty of confidentiality) from a department or directorate
- Other approval E.g. from a Data Protection Officer

Indicate which approval
Upload document (oppdragsdokument)
Chose file...
Upload approvals
Chose file...

Processing

Where will the personal data be processed?
No personal data is processed.

- Computer belonging to the institution responsible for the project
- Mobile device belonging to the data controller
- Physically isolated computer belonging to the data controller
- External service or network
- Private device

Upload guidelines/approval for processing personal data on private devices
Upload

Who will be processing/have access to the collected personal data?

- Project leader
- Student (student project)
- Internal co-workers
- External co-workers/collaborators inside the EU/EEA
- Data processor
- Others with access to the personal data

Which others will have access to the collected personal data? No one.

Will the collected personal data be made available to a third party or international organisation outside the EEA?
No.

Give the name of the institution/organisation
Give the country of the institution/organisation
On what basis will the collected personal data be transferred?
Upload necessary safeguards
Chose file...
Next

Information Security

Will directly identifiable personal data be stored separately from the rest of the collected data (in a scrambling key)?
No.

Explain why directly identifiable personal data will be stored together with the rest of the collected data.

-

Which technical and practical measures will be used to secure the personal data?

- Personal data will be anonymised as soon as no longer needed
- Personal data will be transferred in encrypted form
- Personal data will be stored in encrypted form
- Record of changes
- Multi-factor authentication
- Restricted access
- Access log
- Other security measures
- Indicate which measures

Duration of project

Project period
2019 - 2020

Will personal data be stored beyond the end of project period?
No personal data is processed.

- No, all collected data will be deleted
- No, the collected data will be stored in anonymous form
- Yes, collected personal data will be stored until
- Yes, collected personal data will be stored indefinitely.

For what purpose(s) will the collected personal data be stored?

- Research
- Other

Where will the collected personal data be stored?

- At the institution responsible for the project (data controller)
- Other

Additional information

Will the data subjects be identifiable (directly or indirectly) in the thesis/publications for the project?
No.

Appendix II: Data Management & Analysis Code

```
#setting working directory
setwd("~/OneDrive - Universitetet i Oslo/0.Thesis/R/FinalRfiles")
#install the packages of need
install.packages("readr")
install.packages("readxl")
install.packages("psych")
install.packages("mirt")
install.packages("TAM")
install.packages("ggplot2")
library(readr)
library(readxl)
library(psych)
library(mirt)
library(TAM)
library(ggplot2)


####################################### Data Cleaning #######################
############# Reading test May 2019 #######
#load the response data from reading test from May 2019
may.read <- read.table(file="Les Mai 2019 response data.csv", header=TRUE, sep=";",dec=",")

# subset only the items
may <- may.read[,c(1,3:8,12:19,23:28,32:51,55:74,78:104,108:134,138:168,172:202)]
maydesc <- describe(may)

# 1) cleaning the data with deleting the response number < 10
dat <- as.data.frame(apply(may, 2, function(x) {x[x == 9] <- 99; x})) # gir begge typer omit samme kode lik 99
may[,2:177] <- sapply(may[,2:177], as.numeric)
ut <- dat[rowSums(may[-1], na.rm = T) >= 990, "Kanidat"] #Finner pilotid til de kandidatene radsum lik eller
over 990 som må ha 10 eller flere omit
dat0 <- may[!may$Kanidat %in% ut,]  #data.frame der kanidatene med flere enn 10 omit er fjernet

# rescore 99 and 9 to 0
items1 <- dat0[,2:177]

# deleting the polytonomous item - because there are only binary items in pilot study
items1[1,c(82:87,109:114,134:139,165:170)] # polytomous items
items1 <- items1[,-c(82:87,109:114,134:139,165:170)]

#### making two separate data sets for different analysis
items <- items1

# for calculating the total test length that was presented to each candidate
items1[,][items1[,]==99]=0 # items that are not seen and not completed
items1[,][items1[,]==9]=0  # items that are seen but not completed

# for main data analysis - not including the items that candidates didnt see to the scores
items[,][items[,]==99]=NA # items that are not seen and not completed
items[,][items[,]==9]=0  # items that are seen but not completed

# lukeoppgaver have different names - change them to same name as in the pilot

colnames(items)[which(names(items) == "X1900641")] <- "X1500106"
colnames(items)[which(names(items) == "X1900642")] <- "X1500107"
colnames(items)[which(names(items) == "X1900643")] <- "X1500108"
colnames(items)[which(names(items) == "X1900644")] <- "X1500109"
colnames(items)[which(names(items) == "X1900637")] <- "X1602501"
colnames(items)[which(names(items) == "X1900639")] <- "X1602502"
colnames(items)[which(names(items) == "X1900640")] <- "X1602504"

############################## pilot data ########
# load the data
pilot.read <- read_excel("Final Pilot m 99 response.xlsx")

# delete completed response < 10 for more clean data
data.1 <- as.data.frame(apply(pilot.read, 2, function(x) {x[x == 9] <- 99; x})) # gir begge typer omit samme kode lik 99
pilot.read[,2:302] <- sapply(pilot.read[,2:302], as.numeric)
out <- data.1[rowSums(pilot.read[-1], na.rm = T) >= 990, "pilotid"] #Finner pilotid til de kandidatene radsum lik eller over 990 som må ha 10 eller flere omit
data.2 <- pilot.read[!pilot.read$pilotid %in% out,]  #data.frame der kanidatene med flere enn 10 omit er fjernet

# 4984 - 4846 = 138 observations with less than 10 reponses are deleted.

#rescore pilot data
pilot1 <- data.2[,2:302]

# two different data sets for different analysis
pilot <-pilot1

pilot[,][pilot[,]==99]=NA  # items that are not seen and not completed
pilot[,][pilot[,]==9]=0 # items that are seen but not completed

## test length
tlformal <- items1
tlpilot <- pilot1

# formal data test length for each candidate
respdata <- !is.na(tlformal)
respdata[,] <- as.numeric(respdata[,])
testlength.formal <- as.data.frame(rowSums(respdata))
mean(testlength.formal[,])
median(testlength.formal[,])

# pilot data test length for each candidate
respdata1 <- !is.na(tlpilot)
respdata1[,] <- as.numeric(respdata1[,])
testlength.pilot <- as.data.frame(rowSums(respdata1))
mean(testlength.pilot[,])
median(testlength.pilot[,])

######### Descriptive Statistics ##########
# data frames for descriptive information of items in both data sets
str(items)
summary(items)
str(pilot)
summary(pilot)

descitems <- describe(items)
descpilot <- describe(pilot)

# formal
mean(descitems[1:152,]$n)
median(descitems[1:152,]$n)
mean(descitems[1:152,]$mean)
median(descitems[1:152,]$mean)
```

```
# pilot data set
mean(descpilot[1:301,]$n)
median(descpilot[1:301,]$n)
mean(descpilot[1:301,]$mean)
median(descpilot[1:301,]$mean)

# box plot
Data <- rbind(descitems, descpilot)
Group <- c(rep("Formal", 152), rep("Pilot",301))

colnames(Data)[which(names(Data) == "mean")] <- "P"
colnames(Data)[which(names(Data) == "n")] <- "N"

ggplot(data = Data, aes(x = Group, y = N))+
  stat_boxplot(geom = "errorbar", width = 0.5, na.rm = T) +
  geom_boxplot(stat = "boxplot", outlier.colour = "#ff0000", outlier.size = 1.5, outlier.shape = 8, na.rm = T)

ggplot(data = Data, aes(x = Group, y = P))+
  stat_boxplot(geom = "errorbar", width = 0.5, na.rm = T) +
  geom_boxplot(stat = "boxplot", outlier.colour = "#ff0000", outlier.size = 1.5, outlier.shape = 8, na.rm = T)


############# IRT ###############
### Assumptions
## Unidimensionality
####################### Dimensionality check
# formal
tammod <- tam.mml.2pl(resp = items, irtmodel = "2PL")
Modelfit <- tam.modelfit(tammod)
Modelfit$statlist
Modelfit$Q3_summary

tm3pl <- tam.mml.3pl(resp = items)
tm3plfit <- tam.modelfit(tm3pl)
tm3plfit$Q3_summary

# pilot
tammod1 <- tam.mml.2pl(resp = pilot, irtmodel = "2PL")
Modelfit1 <- tam.modelfit(tammod1)
Modelfit1$statlist
Modelfit1$Q3_summary
Modelfit1$Q3.matr
tam.Q3()

tm3pl1 <- tam.mml.3pl(resp = pilot)
tm3plfit1 <- tam.modelfit(tm3pl1)
tm3plfit1$Q3_summary

# Common items
joint <- merged
str(joint)
summary(joint)
describe(joint)
tammod3 <- tam.mml.2pl(resp = joint, irtmodel = "2PL")
Modelfit3 <- tam.modelfit(tammod3)
Modelfit3$statlist
Modelfit3$Q3_summary

################ DIF
```

```
# find the common items between two datasets
list_df = list(items, pilot)
col_common = colnames(list_df[[1]])
for (i in 2:length(list_df)){
  col_common = intersect(col_common, colnames(list_df[[i]]))
}

# common items are found 56 items
subitems <- subset(items, select=col_common)
subpilot <- subset(pilot, select=col_common)

# Common items in subsets - how do they look like?
describe(subitems)
describe(subpilot)

# Merge the common items in one data frame
merged <- rbind(subitems, subpilot)
stake <- c(rep("formal", 7815), rep("pilot",4846)) # grouping variable


############### 2 step procedure start ###############
# Two steps procedure
step1result <- data.frame(matrix(NA,1,8))
colnames(step1result) <- c("AIC","AICc","SABIC","HQ","BIC","X2","df","p")
for (i in 1:56) {
  # In the first step, we test items, one at a time, by constraining all other items consistent between groups. This
anchor setting is refferred to as "all others as anchors" (AOAA) approach
  testmodel <- multipleGroup(merged, 1, group = stake, SE = TRUE, invariance = c("free_means",
"free_var",colnames(merged[,-i])), method = "EM", technical=list(NCYCLES=10000))
  print.by(cat("DIF analysis is testing Item No.",i, "in the first Step."))
  step1temp <- DIF(testmodel, c('a1', 'd'), items2test = i, technical=list(NCYCLES=10000))
  step1result[i,] <- step1temp
  # we store the result of DIF detection for all items.
}
# We then find out the items which are labeled as "DIF-free item" - Anchor items from step 1
AnchorItems <- step1result[,"p"] > 0.05
colnames(merged[,which(AnchorItems)])

# We treat the DIF-free items in step 1 as the anchor items in step 2. It means that the matching variables in
step2 of DIF testing are the DIF-free items that we got from the result of step 1.
testmodel_step2 <- multipleGroup(merged, 1, group = stake, SE = TRUE, invariance = c("free_means",
"free_var",colnames(merged[,which(AnchorItems)])), method = "EM", technical=list(NCYCLES=10000))
step2result <- DIF(testmodel_step2, c('a1', 'd'), items2test = which(!AnchorItems),
technical=list(NCYCLES=10000))
DIF_items_by2steps <- rownames(step2result[step2result[,"p"] < 0.05,])
# The DIF items is detected by Likelihood ratio test. Yahhhee!!!
print.by(cat("The second step is completed. Please check result in 'step2result'"))
#################### End ########################

# Anchor items from step 2
AnchorItems2 <- rownames(step2result[step2result[,"p"] > 0.05,])
View(AnchorItems2)

# DIF items from step 2
dif2 <- as.data.frame(step2result[step2result[,"p"] < 0.05,])

### group mean difference
coef(testmodel_step2, IRT = TRUE)
tam.fit(testmodel_step2)
coef(testmodel_step2)$pilot$GroupPars
```

```
# New model restricted with final 17 anchor items
totalanchor <- c(colnames(merged[,which(AnchorItems)]),AnchorItems2)
newmg <- multipleGroup(merged, 1, group = stake, invariance = c("free_means", "free_var", totalanchor),
method = "EM", technical=list(NCYCLES=10000), SE = TRUE)
coef(newmg, IRT = TRUE)
coef(newmg)$pilot$GroupPars # report with 95% CI

### Effect size after purification
ES <- empirical_ES(newmg, Theta.focal = NULL,
              focal_items = 1L:extract.mirt(testmodel_step2, "nitems"), DIF = TRUE,
              npts = 61, theta_lim = c(-6, 6), ref.group = 1, plot = FALSE,
              par.strip.text = list(cex = 0.7),
              par.settings = list(strip.background = list(col = "#9ECAE1"),
                          strip.border = list(col = "black")))
ES1 <- ES #for another data set of anonymity later
rownames(ES) <- colnames(merged)

# cohens criteria
medcohensd <- abs(ES[,"ESSD"]) > 0.5
larcohensd <- abs(ES[,"ESSD"]) > 0.8
table(medcohensd)
table(larcohensd)

LargeESmatrix <- ES[larcohensd,]
rownames(LargeESmatrix)[c(1:10)] # DIF items with large ESSD

# change the names for anonymity
larcohensd1 <- abs(ES1[,"ESSD"]) > 0.8
LargeESmatrix1 <- ES1[larcohensd1,]
rownames(LargeESmatrix1) <- sub("item.", "Item ", rownames(LargeESmatrix1))
rownames(LargeESmatrix1) # providing the order of DIF items that can be used in plots with anonymity
# Final DIF items: "Item 2"  "Item 7"  "Item 8"  "Item 16" "Item 26" "Item 48" "Item 51" "Item 53" "Item 54"
"Item 56"

## ICC on final 10 DIF items with large effect size
# item 2
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 1), 2), seq(-8,8, by = 0.01))[,2], type =
"l", xlab = "", ylab = "", ylim = c(0,1), lwd = 2.5, main = "Item 2")
par(new = TRUE)
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 2), 2), seq(-8,8, by = 0.01))[,2], type =
"l", lty=2 ,xlab = "θ", ylab = "Probability",bty='L', ylim = c(0,1), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 7
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 1), 7), seq(-8,8, by = 0.01))[,2], type =
"l", xlab = "", ylab = "", ylim = c(0,1), lwd = 2.5, main = "Item 7")
par(new = TRUE)
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 2), 7), seq(-8,8, by = 0.01))[,2], type =
"l", lty=2 ,xlab = "θ", ylab = "Probability",bty='L', ylim = c(0,1), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 8
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 1), 8), seq(-8,8, by = 0.01))[,2], type =
"l", xlab = "", ylab = "", ylim = c(0,1), lwd = 2.5, main = "Item 8")
par(new = TRUE)
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 2), 8), seq(-8,8, by = 0.01))[,2], type =
"l", lty=2 ,xlab = "θ", ylab = "Probability",bty='L', ylim = c(0,1), lwd = 2.5)
```

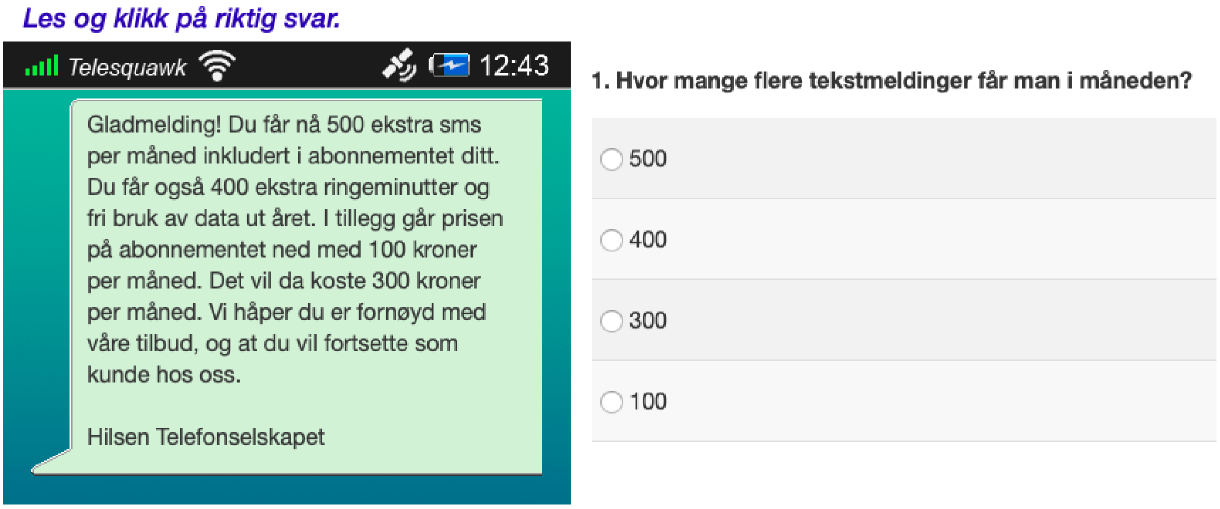
```
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 16
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 1), 16), seq(-8,8, by = 0.01))[,2], type =
"l", xlab = "", ylab = "", ylim = c(0,1), lwd = 2.5, main = "Item 16")
par(new = TRUE)
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 2), 16), seq(-8,8, by = 0.01))[,2], type =
"l", lty=2 ,xlab = "θ", ylab = "Probability",bty='L', ylim = c(0,1), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)


# item 26
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 1), 26), seq(-8,8, by = 0.01))[,2], type =
"l", xlab = "", ylab = "", ylim = c(0,1), lwd = 2.5, main = "Item 26")
par(new = TRUE)
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 2), 26), seq(-8,8, by = 0.01))[,2], type =
"l", lty=2 ,xlab = "θ", ylab = "Probability",bty='L', ylim = c(0,1), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 48
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 1), 48), seq(-8,8, by = 0.01))[,2], type =
"l", xlab = "", ylab = "", ylim = c(0,1), lwd = 2.5, main = "Item 48")
par(new = TRUE)
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 2), 48), seq(-8,8, by = 0.01))[,2], type =
"l", lty=2 ,xlab = "θ", ylab = "Probability",bty='L', ylim = c(0,1), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 51
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 1), 51), seq(-8,8, by = 0.01))[,2], type =
"l", xlab = "", ylab = "", ylim = c(0,1), lwd = 2.5, main = "Item 51")
par(new = TRUE)
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 2), 51), seq(-8,8, by = 0.01))[,2], type =
"l", lty=2 ,xlab = "θ", ylab = "Probability",bty='L', ylim = c(0,1), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 53
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 1), 53), seq(-8,8, by = 0.01))[,2], type =
"l", xlab = "", ylab = "", ylim = c(0,1), lwd = 2.5, main = "Item 53")
par(new = TRUE)
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 2), 53), seq(-8,8, by = 0.01))[,2], type =
"l", lty=2 ,xlab = "θ", ylab = "Probability",bty='L', ylim = c(0,1), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 54
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 1), 54), seq(-8,8, by = 0.01))[,2], type =
"l", xlab = "", ylab = "", ylim = c(0,1), lwd = 2.5, main = "Item 54")
par(new = TRUE)
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 2), 54), seq(-8,8, by = 0.01))[,2], type =
"l", lty=2 ,xlab = "θ", ylab = "Probability",bty='L', ylim = c(0,1), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 56
```

```
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 1), 56), seq(-8,8, by = 0.01))[,2], type =
"l", xlab = "", ylab = "", ylim = c(0,1), lwd = 2.5, main = "Item 56")
par(new = TRUE)
plot(seq(-8,8, by = 0.01), probtrace(extract.item(extract.group(newmg, 2), 56), seq(-8,8, by = 0.01))[,2], type =
"l", lty=2 ,xlab = "θ", ylab = "Probability",bty='L', ylim = c(0,1), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)



###### Item information function

# item 2
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 1), 2), seq(-6,6, by = 0.01)), type = "l",
xlab = "", ylab = "", ylim = c(0,2.5), lwd = 2.5, main = "Item 2")
par(new = TRUE)
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 2), 2), seq(-6,6, by = 0.01)), type = "l",
lty=2 ,xlab = "θ", ylab = "Information",bty='L', ylim = c(0,2.5), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 7
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 1), 7), seq(-6,6, by = 0.01)), type = "l",
xlab = "", ylab = "", ylim = c(0,2.5), lwd = 2.5, main = "Item 7")
par(new = TRUE)
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 2), 7), seq(-6,6, by = 0.01)), type = "l",
lty=2 ,xlab = "θ", ylab = "Information",bty='L', ylim = c(0,2.5), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 8
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 1), 8), seq(-6,6, by = 0.01)), type = "l",
xlab = "", ylab = "", ylim = c(0,2.5), lwd = 2.5, main = "Item 8")
par(new = TRUE)
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 2), 8), seq(-6,6, by = 0.01)), type = "l",
lty=2 ,xlab = "θ", ylab = "Information",bty='L', ylim = c(0,2.5), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 16
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 1), 16), seq(-6,6, by = 0.01)), type = "l",
xlab = "", ylab = "", ylim = c(0,2.5), lwd = 2.5, main = "Item 16")
par(new = TRUE)
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 2), 16), seq(-6,6, by = 0.01)), type = "l",
lty=2 ,xlab = "θ", ylab = "Information",bty='L', ylim = c(0,2.5), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 26
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 1), 26), seq(-6,6, by = 0.01)), type = "l",
xlab = "", ylab = "", ylim = c(0,2.5), lwd = 2.5, main = "Item 26")
par(new = TRUE)
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 2), 26), seq(-6,6, by = 0.01)), type = "l",
lty=2 ,xlab = "θ", ylab = "Information",bty='L', ylim = c(0,2.5), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 48
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 1), 48), seq(-6,6, by = 0.01)), type = "l",
xlab = "", ylab = "", ylim = c(0,2.5), lwd = 2.5, main = "Item 48")
```

```
par(new = TRUE)
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 2), 48), seq(-6,6, by = 0.01)), type = "l",
lty=2 ,xlab = "θ", ylab = "Information",bty='L', ylim = c(0,2.5), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 51
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 1), 51), seq(-6,6, by = 0.01)), type = "l",
xlab = "", ylab = "", ylim = c(0,4.2), lwd = 2.5, main = "Item 51")
par(new = TRUE)
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 2), 51), seq(-6,6, by = 0.01)), type = "l",
lty=2 ,xlab = "θ", ylab = "Information",bty='L', ylim = c(0,4.2), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 53
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 1), 53), seq(-6,6, by = 0.01)), type = "l",
xlab = "", ylab = "", ylim = c(0,2.5), lwd = 2.5, main = "Item 53")
par(new = TRUE)
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 2), 53), seq(-6,6, by = 0.01)), type = "l",
lty=2 ,xlab = "θ", ylab = "Information",bty='L', ylim = c(0,2.5), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 54
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 1), 54), seq(-6,6, by = 0.01)), type = "l",
xlab = "", ylab = "", ylim = c(0,2.5), lwd = 2.5, main = "Item 54")
par(new = TRUE)
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 2), 54), seq(-6,6, by = 0.01)), type = "l",
lty=2 ,xlab = "θ", ylab = "Information",bty='L', ylim = c(0,2.5), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

# item 56
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 1), 56), seq(-6,6, by = 0.01)), type = "l",
xlab = "", ylab = "", ylim = c(0,2.5), lwd = 2.5, main = "Item 56")
par(new = TRUE)
plot(seq(-6,6, by = 0.01), iteminfo(extract.item(extract.group(newmg, 2), 56), seq(-6,6, by = 0.01)), type = "l",
lty=2 ,xlab = "θ", ylab = "Information",bty='L', ylim = c(0,2.5), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)


#### Test characteristic function
plot(seq(-6, 6, by = 0.01), expected.test(extract.group(newmg, 1), matrix(seq(-6, 6, by = 0.01))), type = "l", xlab
= "", ylab = "", ylim = c(0, 60), lwd = 2.5, main = "Test Characteristic Curve")
par(new = TRUE)
plot(seq(-6, 6, by = 0.01), expected.test(extract.group(newmg, 2), matrix(seq(-6, 6, by = 0.01))), type =
"l",lty=2, xlab = "θ", ylab = "Expected Trait Score", ylim = c(0, 60), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)

#### Test information function
plot(seq(-6, 6, by = 0.01), testinfo(extract.group(newmg, 1), seq(-6, 6, by = 0.01)), type = "l", xlab = "", ylab =
"", ylim = c(0, 50), lwd = 2.5, main = "Test Information Function")
par(new = TRUE)
plot(seq(-6, 6, by = 0.01), testinfo(extract.group(newmg, 2), seq(-6, 6, by = 0.01)), type = "l",lty=2, xlab = "θ",
ylab = "Information", ylim = c(0, 50), lwd = 2.5)
par(xpd=TRUE)
legend("topleft", lty= 1:2, legend = c("Formal", "Pilot"), box.lty=1)
```

Appendix III: Examples of Reading Test Items in Norskprøven (Skills Norway, 2019)

1. Multiple-Choice Item (Short Form) in A2/B1 Level:



Les og klikk på riktig svar.

Gladmelding! Du får nå 500 ekstra sms per måned inkludert i abonnementet ditt. Du får også 400 ekstra ringeminutter og fri bruk av data ut året. I tillegg går prisen på abonnementet ned med 100 kroner per måned. Det vil da koste 300 kroner per måned. Vi håper du er fornøyd med våre tilbud, og at du vil fortsette som kunde hos oss.

Hilsen Telefonselskapet

1. Hvor mange flere tekstmeldinger får man i måneden?

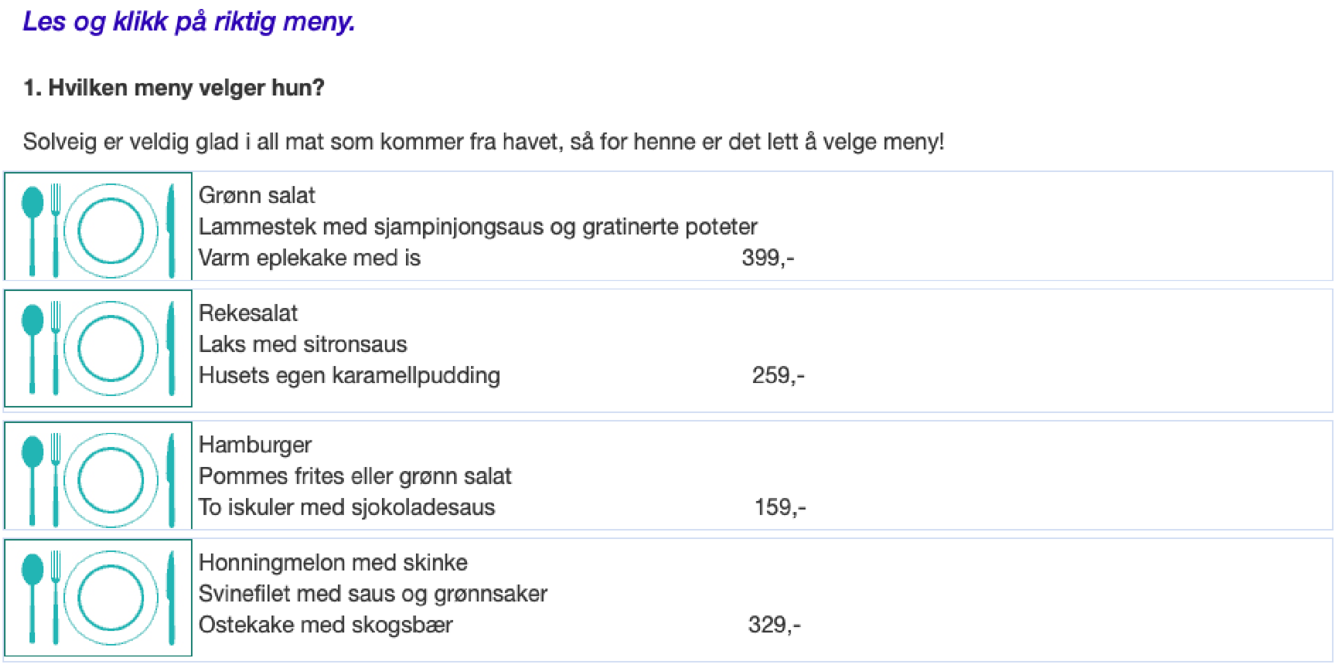
○ 500

○ 400

○ 300

○ 100

2. "Choose Text" Item in A2/B1 Level:



Les og klikk på riktig meny.

1. Hvilken meny velger hun?

Solveig er veldig glad i all mat som kommer fra havet, så for henne er det lett å velge meny!

Grønn salat
Lammestek med sjampinjongsaus og gratinerte poteter
Varm eplekake med is 399,-

Rekesalat
Laks med sitronsaus
Husets egen karamellpudding 259,-

Hamburger
Pommes frites eller grønn salat
To iskuler med sjokoladesaus 159,-

Honningmelon med skinke
Svinefilet med saus og grønnsaker
Ostekake med skogsbær 329,-

Appendix IV: Results Table of 39 DIF Items

DIF Item	AIC	AICc	SABIC	HQ	BIC	X2	df	p
1	-117.95355	-117.827093	-109.4167763	-112.9710632	-103.0609846	121.953548	2	0.000000e+00
2	-51.859616	-51.733161	-43.3228447	-46.8771315	-36.9670529	55.859616	2	7.417400e-13
3	-57.802853	-57.676398	-49.2660816	-52.8203685	-42.9102899	61.802853	2	3.796963e-14
4	-67.815855	-67.689400	-59.2790836	-62.8333705	-52.9232919	71.815855	2	2.220446e-16
5	-67.793854	-67.667399	-59.2570827	-62.8113695	-52.9012909	71.793854	2	2.220446e-16
6	-4.247278	-4.120823	4.2894937	0.7352068	10.6452854	8.247278	2	1.618551e-02
7	-34.133953	-34.007498	-25.5971816	-29.1514685	-19.2413899	38.133953	2	5.239831e-09
8	-14.997994	-14.871538	-6.4612220	-10.0155089	-0.1054303	18.997994	2	7.492696e-05
9	-80.139002	-80.012547	-71.6022302	-75.1565170	-65.2464384	84.139002	2	0.000000e+00
10	-31.428890	-31.302434	-22.8921179	-26.4464047	-16.5363261	35.428890	2	2.026350e-08
11	-6.106704	-5.980248	2.4300680	-1.1242188	8.7858598	10.106704	2	6.387887e-03
12	-12.200843	-12.074388	-3.6640716	-7.2183584	2.6917202	16.200843	2	3.034112e-04
13	-6.222963	-6.096507	2.3138090	-1.2404779	8.6696007	10.222963	2	6.027148e-03
14	-31.128777	-31.002322	-22.5920052	-26.1462920	-16.2362135	35.128777	2	2.354415e-08
15	-16.837376	-16.710921	-8.3006042	-11.8548911	-1.9448125	20.837376	2	2.986904e-05
16	-26.335348	-26.208893	-17.7985765	-21.3528633	-11.4427847	30.335348	2	2.586800e-07
17	-13.395751	-13.269296	-4.8589792	-8.4132661	1.4968125	17.395751	2	1.669401e-04

18	-65.916245	-65.789790	-57.3794733	-60.9337601	-51.0236815	69.916245	2	6.661338e-16
19	-61.167093	-61.040637	-52.6303209	-56.1846078	-46.2745292	65.167093	2	7.105427e-15
20	-8.644597	-8.518142	-0.1078257	-3.6621126	6.2479660	12.644597	2	1.795811e-03
21	-16.053979	-15.927524	-7.5172078	-11.0714946	-1.1614161	20.053979	2	4.419099e-05
22	-102.9361	-102.809741	-94.3994250	-97.9537118	-88.0436332	106.936197	2	0.000000e+00
23	-112.2531	-112.126726	-103.7164101	-107.2706969	-97.3606183	116.253182	2	0.000000e+00
24	-85.842296	-85.715841	-77.3055243	-80.8598112	-70.9497326	89.842296	2	0.000000e+00
25	-62.571658	-62.445203	-54.0348868	-57.5891737	-47.6790951	66.571658	2	3.552714e-15
26	-3.858608	-3.732152	4.6781639	1.1238771	11.0339557	7.858608	2	1.965735e-02
27	-43.332760	-43.206305	-34.7959886	-38.3502755	-28.4401969	47.332760	2	5.270140e-11
28	-40.778219	-40.651764	-32.2414478	-35.7957347	-25.8856561	44.778219	2	1.890311e-10
29	-48.855433	-48.728977	-40.3186610	-43.8729479	-33.9628693	52.855433	2	3.331113e-12
30	-31.666966	-31.540510	-23.1301939	-26.6844808	-16.7744022	35.666966	2	1.798941e-08
31	-48.134996	-48.008540	-39.5982240	-43.1525108	-33.2424322	52.134996	2	4.775624e-12
32	-66.071195	-65.944740	-57.5344238	-61.0887107	-51.1786321	70.071195	2	5.551115e-16
33	-37.088380	-36.961924	-28.5516081	-32.1058949	-22.1958163	41.088380	2	1.196112e-09
34	-138.998106	-138.871651	-130.4613347	-134.0156216	-124.1055430	142.998106	2	0.000000e+00
35	-51.859327	-51.732872	-43.3225555	-46.8768423	-36.9667637	55.859327	2	7.418510e-13

36	-112.098942	-111.972486	-103.5621702	-107.1164570	-97.2063784	116.098942	2	0.000000e+00
37	-84.461412	-84.334957	-75.9246408	-79.4789277	-69.5688491	88.461412	2	0.000000e+00
38	-12.175681	-12.049225	-3.6389091	-7.1931959	2.7168826	16.175681	2	3.072526e-04
39	-85.023865	-84.897410	-76.4870939	-80.0413807	-70.1313021	89.023865	2	0.000000e+00