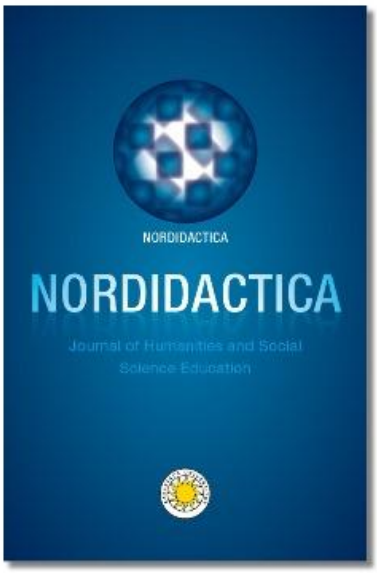
Investigating DIF: The role of response format for students following the syllabus for Swedish as a second language– Experiences from the Swedish National Tests in Social Studies

Arne Löfstedt & Daniel Bergh



Nordidactica

- Journal of Humanities and Social Science Education

2025:2

Nordidactica – Journal of Humanities and Social Science Education
Nordidactica 2025:2
ISSN 2000-9879
The online version of this paper can be found at: www.kau.se/nordidactica

NORDIDACTICA – JOURNAL OF HUMANITIES AND SOCIAL SCIENCE EDUCATION
ISSN 2000-9879
2025.2 73-98

Investigating DIF: The role of response format for students following the syllabus for Swedish as a second language– Experiences from the Swedish National Tests in Social Studies

https://doi.org/10.62902/nordidactica.v15i2025:2.26762

Arne Löfstedt & Daniel Bergh

Göteborgs universitet, Institutionen för pedagogik och specialpedagogik

Abstract: In developing the national tests in Social Studies, a critical issue is ensuring that the items work in the same way for subgroups of students, i.e., the test is fair and shows no signs of severe Differential Item Functioning (DIF). In Sweden, the national tests will be digitalized, but items with a selected response format (SR) is also intended to increase. This study aims to analyze how different items (defined by varying formats and contents) function for students following the syllabus of Swedish as a second language (SvA) compared to the general student population. The study is based on data from the Swedish national tests in Social Studies in 2018, 2019, 2022, and 2023. The results reveal that there are DIF between the two syllabus groups: Students following the SvA syllabus have more difficulties compared to Sv students, with SR items. However, SvA students perform significantly better on some constructed response items where they are asked to describe, analyze, and elaborate on a topic using their own language. As this study shows that the SvA group has significantly more difficulties compared to the general student group with the SR items that are intended to increase due to digitalization, there seems to be an obvious risk for this group to be disadvantaged.

KEYWORDS: NATIONAL TESTS, SOCIAL STUDIES, PSYCHOMETRIC ANALYSIS, DIFFERENTIAL ITEM FUNCTIONING, SWEDISH AS A SECOND LANGUAGE, RESPONSE FORMATS

About the authors: Arne Löfstedt`s research focuses on improving the development of various items in the national tests in Social Studies using CTT and IRT. He has also written about interrater reliability. He has worked with the National Agency for Education for many years to create different syllabi in Social Studies or the compulsory and upper secondary in Sweden.

Daniel Bergh`s research stands on two legs: Psychometric analyses based on Rasch Measurement Theory used in order to develop new instruments, but also to evaluate the psychometric properties of already existing scales. He also conducts research on the associations between school achievements and mental health, investigating moderating factors useful to improve school achievements as well as the mental health of students.

Introduction

In testing and assessment, it is essential to ensure measurement invariance across subgroups of pupils, i.e., that the instrument works similarly for different groups to be compared. This invariance property is also a basic requirement for quantitative comparison. Thus, in order to be able to statistically compare groups of students using an instrument, it is crucial that the instrument work in the same way for the groups that are to be compared.

Given that the national test shall provide a basis for the assessment of pupils' knowledge in relation to subject requirements and support fair assessment and grading, as well as being useful in follow-up evaluations, the property of invariance needs to be an integral part of test development and construction. Provided that it is commissioned that the test should be designed in such a way that it does not disadvantage any group, it is clear that the requirement of the invariance property of the test must be satisfied. Therefore, it is fundamental and necessary to ensure that the test has the same measurement features, and thus works in the same way for different groups of students, i.e., show no Differential Item Functioning (DIF) (Andrich & Maris, 2019); Wu, Tam, & Jen, 2016). The potential consequences of the test on students' life opportunities and chances can be crucial, given that the test is a high stakes test, and as the test results are heavily influential on the final grades, but the test results are also indirectly influencing student opportunities for Upper Secondary school admission.

The Swedish National Agency for Education intend to increase the proportion of items with a selected response format to facilitate automatic assessment. However, the implications of this change have not yet been examined thoroughly.

Therefore, the purpose of this study is to analyze the impact different formats can have on various subgroups of the Swedish national tests in Social Studies[1], using data from four years of investigation (2018, 2019, 2022, and 2023). Analyses presented in the article are conducted by means of modern test theory and more specifically using the polytomous Rasch Model (Rasch, 1960/1980; Andrich & Marais, 2019). However, at different stages of the developmental process also Item Response Theory models including more than one parameter (2-PL) has been used. A specific focus is directed toward the analysis of Differential Item Functioning (DIF). The national tests are

[1]In this article, the term "social studies" is used as a name for the Swedish subject "samhällskunskap" (direct translation: Society Knowledge). The term Civics is used by the Swedish National Agency for Education to describe the subject in the Swedish compulsory school. However, Civics may be considered too narrow compared to Social Studies, as it predominantly encompasses the roles and responsibilities of citizens, along with the political and governmental aspects of citizenship. Thus, Civics focuses on knowledge of politics and democracy. However, the concept of Social Studies may be viewed as too broad to accurately reflect the meaning of the subject "samhällskunskap." A solution to this problem might have been to translate the Swedish subject "samhällskunskap" into the English term "Society-Knowledge," but then we would risk some readers not following what is meant. Therefore, we have consistently chosen to use the term Social Studies throughout this work.

Link to curriculum Lgr22:
Läroplan för grundskolan samt för förskoleklassen och fritidshemmet

reported with two subgroups: sex, and Swedish language syllabus group: Swedish as a second language (SvA) or Swedish (Sv). The pupils are following two different syllabus in Swedish language: Swedish as a second language or those who follow the syllabus of (ordinary) Swedish. Both groups are taking the same course and national test in Social Studies.

As it is very important to be assured that the national test in Social Studies work invariantly across student groups, and since language proficiency levels is a common source of bias in measurement, influencing student performances, it is of particular importance to examine how the instrument works across language proficiency groups. Thus, it is very important to assure that students are not unfairly disadvantaged or advantaged due to their language proficiency levels. Therefore, more specifically, this study aims to respond the following research questions:

- Do selected response (SR) items included in the Swedish national test in Social Studies work invariantly across Swedish language syllabus groups, i.e. show no Differential Item Functioning (DIF)?
- Do Constructed Response (CR) items included in the Swedish national test in Social Studies work invariantly across Swedish language syllabus groups, i.e. show no Differential Item Functioning (DIF)?
- If SR items and CR items do not work invariantly across Swedish language syllabus groups, how is DIF manifested?

Test developers routinely conduct Differential Item Functioning (DIF) analyses during testing periods. However, as noted in a comment below, these details are confidential as they may include reports of precise data. As part of their development activities, the group also engages in research concerning materials (e.g. for Social Studies: Andersson & Larsson, 2023). This study constitutes a component of that development work. Separate research endeavours are consistently necessary, allowing to produce research with broader, more generalizable content. All groups involved in national testing across five Swedish universities conduct research of a similar nature to this study.

The debate on whether different formats can accurately represent a student's understanding of the material is also connected to the subject itself, how that particular subject aligns with its assessment, and how the assessment works for different subgroups. Additionally, one must understand the specific demands that each subject imposes on its assessment, if one wants to investigate how different formats give students a fair opportunity to show their knowledge of the subject. The syllabus in the subject of Social Studies puts special demands on understanding the content, reflecting modern society. Thus, to understand the test, you must understand the subject.

Background

The subject

Social Studies is a mandatory subject in Swedish compulsory schools as well as in upper secondary schools, and it has a 60-year history in Sweden as a single subject. Thus, it is one of the youngest subjects in the Swedish school system. Given its structure and focus, Social Studies exist as a compulsory subject only in the Nordic countries and has slightly different content and status in Sweden compared to the other Nordic countries (Christiansen 2015a; Gulberg 2023; Spanget Christensen 2023; Børhaug 2023). In Sweden, Social Studies hold a strong position and is regarded as a useful and interesting subject by students. From a teacher's perspective, Social Studies is described as a subject that is fun, yet demanding to teach (Skolverket 2005).

The areas of 'democracy' and 'politics' are two, among several, core areas of Social Studies. In the latest ICCS (International Civic and Citizenship Education Study), Swedish 14-year-olds performed well on politics and democracy tasks (ICCS 2022).

Brief history and structure of the subject

Social Studies in primary and lower secondary schools are today a part of the subject group social study subjects[2], which also includes History, Geography, and Religious Education. Up to and including school year 6, these are often taught together as a subject block. History, Geography, and Religious Education were already defined as individual subjects in the public school[3], laying the foundation for the Swedish comprehensive school in the 19th century. However, that was not true also as regards Social Studies. Especially in History, there have been some early elements regarding the current structure of society. It was only after the 1918 reform on universal suffrage that the idea of preparing students for participation in the democratic process emerged (Larsson & Ledman 2023). The focus was then more on 'civic knowledge.' The subject of History was changed in 1928 to 'History with Social Studies.' (Larsson & Ledman 2023).

Already before the Second World War, a major reform was discussed regarding a comprehensive school to replace the educational system in Sweden, which had parallel systems divided by gender and class. During the 1950s, extensive experimentation occurred with a unified school, and a political compromise in 1957 laid the foundations for the proposal for a unified comprehensive school[4], which was implemented in 1962 (Lärarnas historia 2024).

Within the framework of the comprehensive school, Social Studies would become an individual subject, focusing mainly on democratic education, social life, and institutions in Sweden and internationally (Larsson & Ledman 2023). Modern society

[2] Social study subjects is the translation of the subject group “samhällsorienterande ämnen”, abbreviated into SO, according to the translation provided by the Swedish National Agency for Education.
[3] Translation of the Swedish concept of “Folkskolan”.
[4] Translation of the Swedish concept of "Grundskolan".

demanded many new areas to be included in a new subject. The "moral education" that had previously been dealt with in the subject of Christianity should become more of a "civic education" from a modern perspective. All this was assigned to the new subject of Social Studies (Landahl 2015). Even before the new curriculum in compulsory school, Lgr 62, Social Studies was partly taught as an independent subject, but with the syllabus in Lgr 62, its status as an individual and separate subject had been established (Lgr69).

From the implementation of compulsory education until today, there have been six curricula: Lgr 62, Lgr 69, Lgr 80, Lpo 94, Lgr 11, and Lgr 22. The number indicates the year when the curriculum came into force.

During the subsequent revisions of the curricula (Lgr 80, Lpo 94, and Lgr 11), the subject was developed to include greater elements of reflection and reasoning on national and international social issues. Areas addressing sociological topics, such as the conditions for children's upbringing, were also included.

From Lgr 11, there is a shift towards, in addition to reflection on social problems and issues, a focus on "factual knowledge" within the subject. The move towards more facts and fewer central themes continues with the syllabus in Lgr22.

To examine the development of the subject content, one can also investigate the requirements for Social Studies teachers, i.e., teachers in lower and upper secondary schools. Although Social Studies hold a strong position in the school system, it has never been a separate academic discipline at universities in Sweden.

Before compulsory education, the inclusion of academic disciplines in social studies teacher degrees was debated. Various proposals emerged for upper and lower secondary schools but were not deemed economically viable (Eklund & Larsson 2009). From 1962 to 2001, degree content varied across Swedish universities; however, upper secondary school teaching required three semesters while lower secondary school required two (Eklund & Larsson 2009). Political science and economics were consistently included in Social Studies. Over the years, the focus shifted between social dimensions and business-related topics. Content evolution affected perspectives on academic education. In the 1960s, only political science and economics were required; by the 1970s and 1980s, four subjects—political science, economics, human geography, and sociology—were needed. However, the multidisciplinary content of the subject becomes a problem in an educational context:

> *Because the subject known as Social Studies in schools corresponds to a number of different subjects in universities, a teacher training program comprising three or four grades must necessarily be incomplete, regardless of how the grades are combined* (translated from Brolin 1965, in Larsson 2009, s. 215).

The curricula and assessment in the subject

Clear descriptions of how a teacher can carry out summative assessments[5] of individual students' performance as a basis for grades are a relatively new phenomenon in secondary Social Studies curricula and syllabi. Nevertheless, teachers have been marking individual students in the subject since 1962.

A brief summary of the features and opportunities for teachers to receive help with individual assessment is provided. It should be noted that Sweden does not have a system of national tests that directly determines grades in subjects. Teachers in Sweden have a lot of freedom to grade individual students (Lundahl 2017).

From the perspective of how the curricula and syllabuses have described "assessment of individual pupils," it can be said that the first three syllabuses had no content on this at all. Examination is carried out by identifying a very large number of subject areas. In Lgr69, this consisted of around 180 subject areas. From Lpo94 and onwards, there is a clear shift, even though the texts in Lpo94 are more general regarding how to assess an individual student. From Lgr11 and now Lgr22, several pages in the syllabus describe how an individual student can demonstrate knowledge of the subject.

National tests in Social Studies

Since the 1960s, there have been variants of national tests in Swedish, English, and Mathematics. The focus of the tests has shifted over the years. It was not until 2009 that the subjects were expanded to include national tests for year 9 in science studies (Physics, Biology, and Chemistry[6]). In 2013, national tests in the social study subjects (Social Studies, Geography, History and Religious Education) were introduced for the first time.

At the same time, year 6 tests were conducted in these subjects and in the science studies subjects. The year 6 science and Social Studies tests were abolished in 2015.

The Social Studies test is taken at the end of the second semester (of two) in year 9, over two days. The test grade levels are determined in advance; thus, no standardization or normalization is conducted afterwards. The Social Studies test has been administered 10 times. In 2020 and 2021, no compulsory exams were given due to the COVID-19 pandemic. Despite the multidisciplinary approach of the subject, surveys have demonstrated good assessment consistency and inter-rater reliability of the test (Löfstedt 2018).

As already mentioned, teachers in Sweden have great freedom to set grades. Since 2018, a statement from the Swedish National Agency for Education has been in force, whereby teachers must "take particular account" of the test grade from the national tests when grading their students (SKOLFS 2022:417). Research has indicated difficulties for teachers in interpreting this concept (Sandberg 2021).

---

[5] "The purpose of summative assessment is to evaluate student learning at the end of a unit of instruction by comparing it to some standard or benchmark." Carnegie Mellon University.
[6] Science studies is the translation of the subject group "naturorienterande ämnen", abbreviated into NO, according to the translation provided by the Swedish National Agency for Education.

As of 2022, the purpose of the national test is now solely "to support grading" (SKOL-FS 2022:185, 9 kap 21 §). Previously, there was also the purpose of providing information at an aggregate level on how the objectives are met and of concretizing the curricula (Hirsch 2015).

Methods

Participants and data collections

This study is based on data from students who took the Swedish national tests in Social Studies in school year 9, included in the cohorts taking the test the following years: 2018, 2019, 2022, and 2023, which represent the last four years of data collection when this study was carried out. During the years 2020-2021, no data collection took place due to the COVID-19 pandemic. Each cohort represents a whole age group corresponding to approximately 100,000-120,000 students. The data consists of all tasks included in the test each year. Given that the Social Studies test is one out of four tests among the social study subjects (Geography, History, Religious Education, Social Studies), 25 percent of the total age cohort take this particular test.

Procedure

All teachers whose students are taking the tests in Social Studies all around Sweden, are invited by the group at University of Gothenburg that constructs the tests, to report the results on a digital reporting service. The teachers report pupils born on the 6th, 16th, and 26th of each month. Teachers report two subgroups: sex (boy, girl) and syllabus followed in Swedish language (Swedish (Sv) or Swedish as a second language (SvA)). Both groups take the same test; they are just subgroups in the total set of pupils. The following applies to who may study under the syllabus for Swedish as a second language:

It is the headmaster who decides whether a pupil in compulsory school or an equivalent type of school should receive teaching in Swedish as a second language.

The decision must be based on the pupils' needs, but it is not regulated in detail how the need assessment is to be carried out (SFS 2011:185, 5 kap 14 §).

Teaching in Swedish as a second language shall, if necessary, be organized for:

1. pupils who have a language other than Swedish as their mother tongue,

2. pupils who have Swedish as their mother tongue and who have been admitted from schools abroad, and

3. immigrant pupils whose main language of interaction with a guardian is Swedish.

It is thus not a subject aimed at all pupils who need support in their language development (Nationellt för centrum för svenska som andraspråk).

In this study, the total number of pupils in the sample for the four tests was 6,759, of whom 639 (9.3%) were following SvA. As shown in Table 1, the proportions of students

following the syllabus for Swedish as a second language have varied somewhat during the four years of investigation, with 11.8 percent at the highest and 7.5 percent at the lowest.

TABLE 1

The number of students participating each year of investigation, in total 6759, divided by the syllabus followed in Swedish language. In addition, information on the number of tasks and the distribution of selected response tasks and constructed response tasks is provided.

Year	Number of students in each year sample	Number of students following SvA	Number of tasks in each test	Distribution of selected response/constructed response tasks
2018	1451	171	24	12/12
2019	1509	134	24	14/10
2022	1987	149	24	15/9
2023	1812	175	29	18/11

From Table 1, it is also notable that the number of tasks is increasing over time. Thus, the number of tasks is greater in 2023 (29) compared to 2018 (24). In addition, the number of tasks with a selected response format has increased over time. Therefore, the proportion of tasks with a selected response format is highest in 2023 (62 percent), while the proportion of tasks using a constructive response format is lowest at 38 percent. These changes align with the intentions of the Swedish National Agency for Education and the Swedish Ministry of Education to increase the number of tasks that can be assessed automatically, which require a selected response format (Utbildningsdepartementet, 2017).

The national tests in Social Studies are developed by the Group for the National Tests in Social Studies at the University of Gothenburg on behalf of the Swedish National Agency for Education. Each year, extensive psychometric analyses of the Social Studies tests are carried out. Detailed psychometric analyses of tasks and the included items are also conducted before the tests, and the tasks included in each test, are subjected to psychometric evaluation and testing 2-3 years before the test is administered. This study is based on the complete data reported by teachers into the system. Therefore, the data is not collected in order to satisfy the aim of this study. Thus, this study use data already collected within the national tests of Social Studies. The data are analysed using methods within the Modern Test Theory paradigm, particularly the Polytomous Rasch Model, with a specific focus on Differential Item Functioning (DIF), to determine whether the test and the tasks function invariantly across different syllabus groups, i.e. that the items work in same way for subgroups of students. The methods used are briefly described below.

Psychometric Analyses

In this study, the concordance between observed data and the expected Rasch Model (Rasch, 1960/1980) is analysed using the Rasch model for ordered response categories, also known as the Polytomous Rasch Model. The polytomous Rasch Model is an extension of the Simple Logistic Model which can be defined as follows:

$$\Pr\{X_{ni} = x\} = \frac{e^{x\beta_n - \delta_i)}}{1 + e^{\beta_n - \delta_i}}.$$

Thus, in the dichotomous case, item locations (item difficulties) are denoted by δ, and person locations (person abilities) are denoted by β. The interactions between items and persons are crucial, and the relationship can be described as follows: the probability of a specific response is a function of the relationship between person parameter (ability) estimates and item parameter estimates (difficulty), resulting in β - δ. A positive value from the subtraction implies probabilities greater than 0.5; accordingly, the person is more likely to respond correctly to the item than incorrectly. On the contrary, if a person has a lower ability than the item is difficult, the person is less likely to respond correctly than incorrectly to the item. Commonly, social scientific data are not restricted to dichotomous response formats. Instead, the polytomous response format may be more applicable in many situations. The Polytomous Rasch Model, or the Rasch model for ordered response categories (Andrich, 1978; Wright & Masters, 1982), takes the general form:

$$\Pr\{x_{ni} = x\} = \frac{e^{-\tau_{1i} - \tau_{2i} \ldots - \tau_{xi} + x(\beta_n - \delta_i)}}{\sum_{x'=0}^{m_i} e^{-\tau_{1i} - \tau_{2i} \ldots - \tau_{x'i} + x'(\beta_n - \delta_i)}}$$

Thus, a central concept needs to be introduced with a polytomous response format: *threshold*. Given a situation with four response categories (0, 1, 2, 3), a threshold specifies the point (or location) at which the probability of choosing one out of two answers is equal, for instance, an answer of 0 or 1 in the example above. Analogously, in an achievement test, the threshold between 0 and 1 specifies the point between Fail and Pass, and the threshold between 1 and 2 the point between Pass and Pass with Distinction. The threshold concept is also important since this is where most information is found. In the equation above, the threshold parameter is denoted by τ and the item score by x in the numerator. Given that there is concordance between the expected Rasch model and the data, the item discriminations (slopes) in the intersections between response categories, i.e. at the thresholds, are the same, as illustrated in Figure 1.

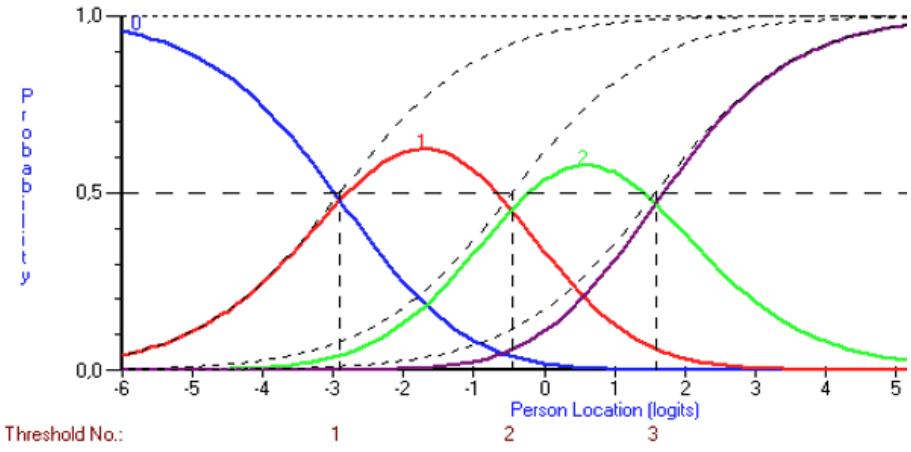


FIGURE 1

An example of a category probability curve showing latent dichotomous threshold characteristic curves with equal slopes.

If the data fit the Rasch Model, the raw score (the summation of item responses) constitutes a sufficient statistic. This means that no additional information can be extracted from the data beyond what the raw score provides. Since the measurement invariance property is an integral part of the Rasch Model, it is possible to test whether the items work invariantly across different classifications of individuals who are to be compared statistically (Andrich, 1988). Thus, if the operational characteristics are similar for the items included in the test, i.e. if the items work invariantly (for different classifications of students) and the data thus conform to the Rasch model, using person measures based on the summation of raw scores across items is justified. However, there are situations when not all the requirements of the Rasch Model can be satisfied. One such requirement regards the assumption of equal discriminations across items. If the requirement for equal discrimination cannot be satisfied, this requirement can be relaxed by introducing the discrimination parameter into the model, as Birnbaum (1968) suggested. Thus, the discrimination parameter characterizes the slope of the Item Characteristic Curve (Andrich & Marais, 2019):

$$\Pr\{X_{ni} = 1|\beta_n, \delta_i, \alpha_i\} = e^{\alpha_i(\beta_n - \delta_i)}/\left[1 + e^{\alpha_i(\beta_n - \delta_i)}\right],$$

In the equation above, the person's ability for person n is denoted by β_n, and the difficulty of the item by δ_i, followed by that item's discrimination, α_i. Given the complexity of the subject under study here, it is reasonable to assume that the requirement of equal discriminations across items cannot be assumed. Therefore, the extension suggested by Birnbaum (1968) including also a discrimination parameter,

relaxing the requirement of equal discriminations, has been implemented at different stages of the development process proceeding this study.

However, it may be relevant to clarify for the interested reader that the Rasch Measurement Theory (RMT) paradigm and the Item Response Theory (IRT) paradigm have a lot in common, given only one parameter models are used, as they have similar parametrization. Nevertheless, they come from different traditions. RMT researchers and IRT researchers differ not least in their view of the relationship between the data and the model, but also purposes of analyses, as is illustrated by Andrich (2011). Thus, according to Andrich (2011) the dominant criterion for model choice within the IRT approach is to find a model that fit data the best, whereas within the Rasch Measurement approach the model is specified a priory with specific focus to improve the instrument and measurement. Therefore, in this sense, in the Rasch tradition the model is considered to be fixed, and the researcher examines whether the data fits this model. Within the IRT tradition, the model is flexible and is adapted in order to find fit to the data. The results presented in this article are based on analyses using the Rasch Model, but within the development process also models applying more than one parameter have been used.

Analysis of Fit

Within the area of modern test theory, it is common to use residual-based fit indices to study the concordance between data and the model. This study is no exception, even though we are not specifically focusing on individual item fit in the presentations. However, general fit statistics (Chi-Squared and their transformations to mean squares (infit/outfit), as well as their graphical representations (ICC)) are used to evaluate the fit to the polytomous Rasch model. The Chi-Square family of statistics for analyzing fit is conducted by comparing the total scores of individuals in approximately equal-sized class intervals, C, with the sum of expected values. Thus, this results in an approximate Chi-Squared statistic with C-1 degrees of freedom. The fit statistics are based on comparisons between observed and expected values. The summation of Chi-Squared values for individual items forms a total Chi-Squared value describing a global overall test of fit for the model. These Chi-Squared statistics can easily be transformed to mean squares by dividing the Chi-Squared value by the degrees of freedom, resulting in the mean square with an expected value of 1 (Wu, Tam & Jen, 2016). As infit and outfit mean squares are commonly used in many psychometric software applications, this transformation is conducted to facilitate communication.

Differential Item Functioning

To statistically test whether items work invariantly across subgroups of individuals, i.e. in the same way for the groups, two different methods were used, in addition to their graphical representations (ICC). Following Wu, Tam and Jen (2016), and Andrich and Marais (2019), Differential Item Functioning is defined to occur: when two groups of people with the same ability level have different probabilities to pass an item, and thus resulting in residual differences between the groups on average. However, it may be

reasonable to point out that analyses of Differential Item Functioning is not limited to modern theory but is also relevant within classical test theory models. One of the most commonly used method within modern test theory, the Mantel-Haenszel test, was applied. The Mantel-Haenszel test examines how stable the strength of the relationship between two binary factors (e.g., syllabus followed, Sv/SvA) is by means of odds ratios across different strata that constitute the levels of the confounding variable (Wu, Tam, & Jen, 2016). Thus, similar odds (odds ratio of 1) indicate no DIF-effect while an odds ratio greater than 1 is favouring the reference group, and an odds ratio less than 1 favours the focal group. However, the Mantel-Haenszel method has been applied to study whether the likelihood of responding to an item correctly is associated with the group to which the individual belongs while controlling for their ability by contrasting a reference group with the focal group (Wu, Tam, & Jen, 2016). The Mantel-Haenszel test is a chi-squared test with one degree of freedom and can be computed by:

$$\chi^2_{MH} = \frac{\left[\left|\sum_k A_{ik} - \sum_k E(A_{ik})\right| - 0.5\right]^2}{\sum_k \mathrm{var}(A_{ik})}$$

In addition to the Mantel-Haenszel test, the occurrence of Differential Item Functioning is identified through statistical analyses of the standardized residuals indicating the distance between the expected model and the observed scores. Thus, the standardized residual for each person is given by:

$$z_{ni} = \frac{x_{ni} - E[x_{ni}]}{\sqrt{V[x_{ni}]}},$$

where $E[x_{ni}]$ is the expected value given person n's and item i's parameter estimates, and $V[x_{ni}]$ is the variance. For the detailed analysis, each person is further identified by their group membership g and class interval c.

This gives the residuals:

$$z_{n_{cg}i} = \frac{x_{n_{cg}i} - E[x_{n_{cg}i}]}{\sqrt{V[x_{n_{cg}i}]}}.$$

The residuals are analyzed by means of analysis of variance (ANOVA). ANOVA is used to determine whether there is a significant difference between the means of two or more groups. In the context of DIF, ANOVA is used to determine whether there is a significant difference among the mean residuals for the groups of interest (Andrich & Marais, 2019). In the context of this study, the differences between students following the syllabus for Swedish (Sv) or Swedish as a second language (SvA), focusing on the probability for the two groups to correctly respond to an item, given the same ability (knowledge based on the result on the complete test, as is understood within the framework usually applied in modern test theory) is analyzed. The statistical analyses of DIF presented in this paper are based on this method described above, using ANOVA analyses of residuals between the groups of interest (Sv and SvA syllabus groups). Significant results from these analyses guided the authors to reject the null hypothesis suggesting that there are no differences between the residuals for Sv and SvA students.

Non-significant results, on the contrary, lead the authors to conclude a situation with no differences, and thus acceptance of the null hypothesis.

Software used

Analyses are conducted using the RUMM2030 plus software (Andrich, Sheridan, & Luo, 2019) and the jMetrik software (Meyer, 2014).

Results

Targeting

Figure 2A – 2D shows the person distributions relative to the distributions of item parameters for the different years. Thus, it is evident that the four tests are well-targeted, i.e., there are item thresholds covering the entire latent trait and various ability levels, meaning there are items of appropriate difficulty for students with low achievements as well as for those with high achievements. It is also clear that most item thresholds align with the locations of the majority of individuals. From this, it follows that the reliability is relatively high. For instance, the Cronbach Alpha (Cronbach, 1951) analogous, the Person Separation Index (Andrich, 1982), sometimes also called the reliability of case estimates (Wright & Masters, 1982), ranges from 0.93 to 0.95, indicating that the instruments can effectively separate individuals with different abilities. This also means that, given this high level of consistency, two parallel investigations would correlate at 0.93 to 0.95, and a very small proportion of the variance can be attributed to factors other than the variations in the included items.

Figure 2A. Person and item distributions 2018

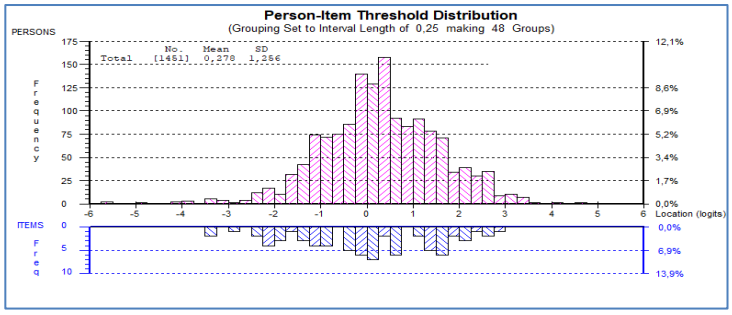


Figure 2B. Person and item distributions 2019

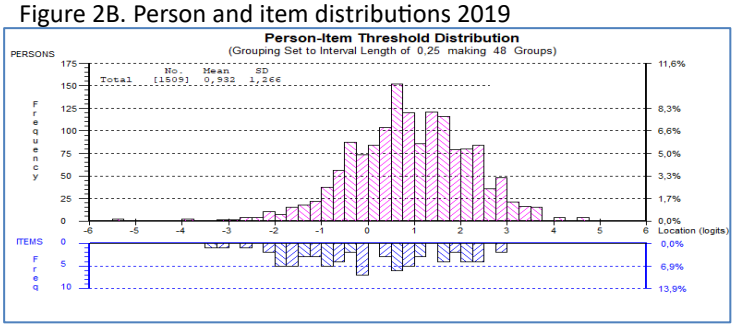


Figure 2C. Person and item distributions 2022

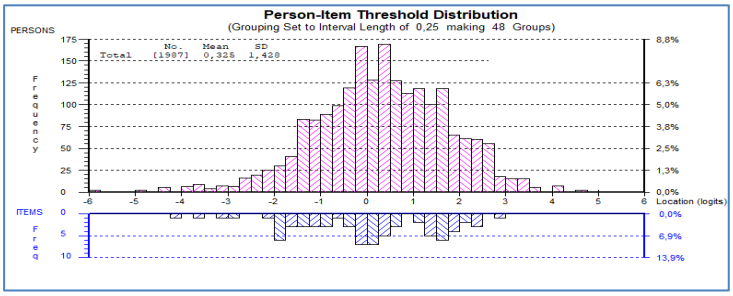


Figure 2D. Person and item distributions 2023

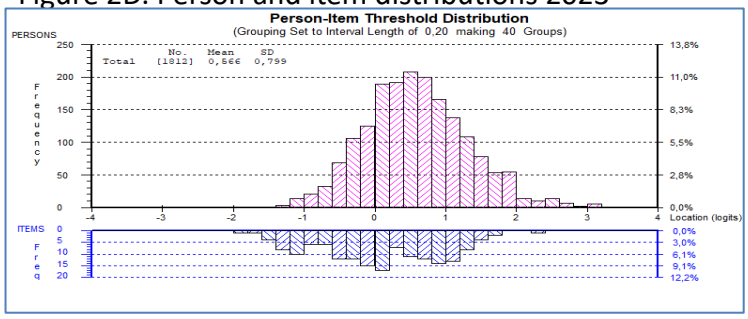


FIGURE 2A-2D

The person distributions relative to the distributions of item parameters for the different years (2018, 2019, 2022, and 2023).

Table 2 shows the mean locations divided by the syllabus followed in each year of investigation. Thus, the general pattern is that students following the Swedish syllabus achieve a higher total score than those following the syllabus for Swedish as a second language. These differences are consistent for all years of investigation. It is noticeable that students following the syllabus for Swedish as a second language have a negative mean location for all years of investigation; moreover, the differences between these students and those following the syllabus for Swedish are more than one logit for all years except 2023.

TABLE 2

Mean locations (Standard Deviation) along four tests, divided by syllabus followed

	2018	**2019**	**2022**	**2023**
Syllabus				
Sv	0.410 (1.19)	1.050 (1.20)	0.482 (1.35)	0.631 (0.79)
SvA	-0.712 (1.33)	-0.281 (1.32)	-1.010 (1.39)	-0.036 (0.67)

Results from analysis of fit – two examples

Figure 3A shows an example of an item (all figures of items are from tasks in the actual tests) that fits well with the expected model. The graph displays the performance of an individual item relative to the performance of pupils with different achievement levels. A low (negative) value implies low achievement on the test, while a high (positive) value implies high achievement. It is also evident that increasing levels of achievement increase the likelihood of the individual obtaining a higher score on this particular item. The item works well with other items, as the observations (dots) lie on the expected curve (line), which is also confirmed by the Chi-Square value of 1.64, which is clearly non-significant, given 5 degrees of freedom (P=0.89). The insignificant Chi-Square value implies that the deviation between the operational characteristic of this particular item and the expected model (i.e. relative to the characteristics of other items) is small, and thus the null hypothesis (no differences) needs to be accepted. On the contrary, Figure 3B shows an example of an item that somewhat misfits to the expected model, i.e. an item that works differently compared to other items, and the expected model. Thus, persons with lower achievement obtain a higher score on this particular item than expected by the model, while students with higher achievement obtain a score lower than expected by the model. Consequently, the Chi-Square value indicates misfit (X^2 = 64.01, d.f.=5, P<0.0001), and thus the null hypothesis (no differences) needs to be rejected. The corresponding Mean-Square value (with the expected value of 1) of 12.8 also confirms this.

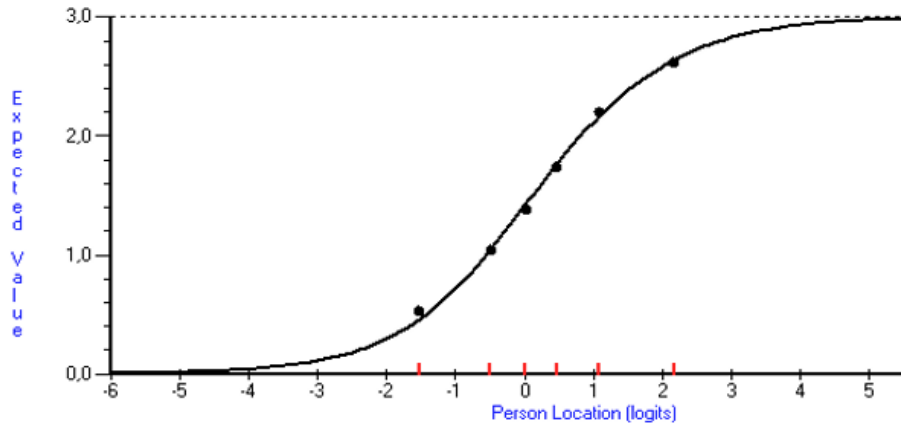


FIGURE 3A

Example of a fitting item

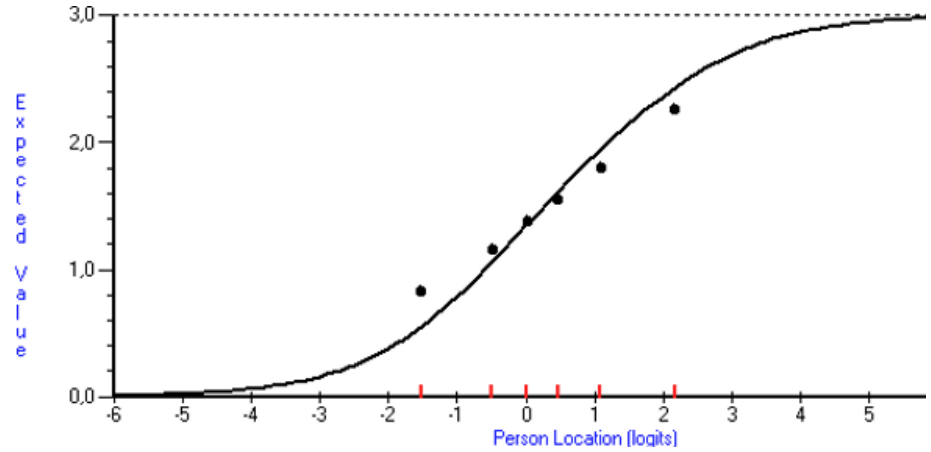


FIGURE 3 B

Example of misfitting item

Results from analyses of Differential Item Functioning (DIF)

In Figure 4, the graphical representation of the analysis of DIF is presented. The analysis shows a situation where no significant DIF effects can be observed, which is also reflected in the statistical test of DIF using ANOVA of residuals (F = 0.34; d.f.=1; P = 0.88). The item works in the same way for the pupils following the syllabus for Swedish (blue) as well as for those following the Swedish as a second language syllabus (red). The comparison is the same along the entire latent trait, i.e., among low achievers as well as among high achievers. Thus, the null hypothesis (no differences between group means) is accepted, and it is concluded that the item works invariantly across the syllabus groups.

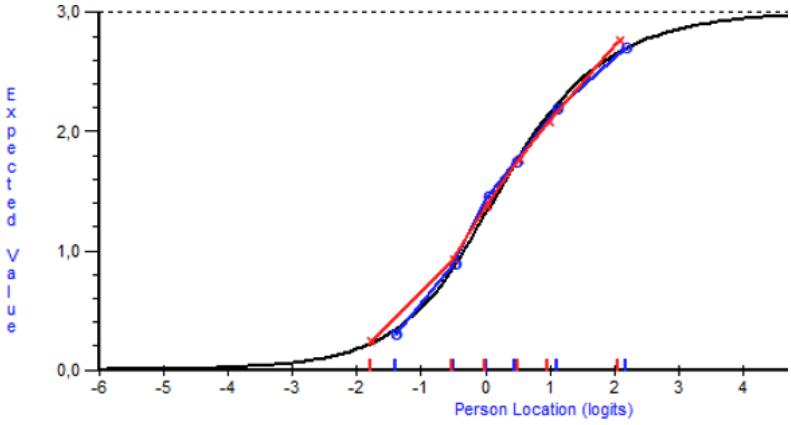


FIGURE 4

Graphical display of analysis of Differential Item Functioning (DIF) with no DIF effect present (SV represented by blue line, SvA red line).

However, it is not always the case that the items work invariantly across the syllabus groups. Figure 5 gives such an example.

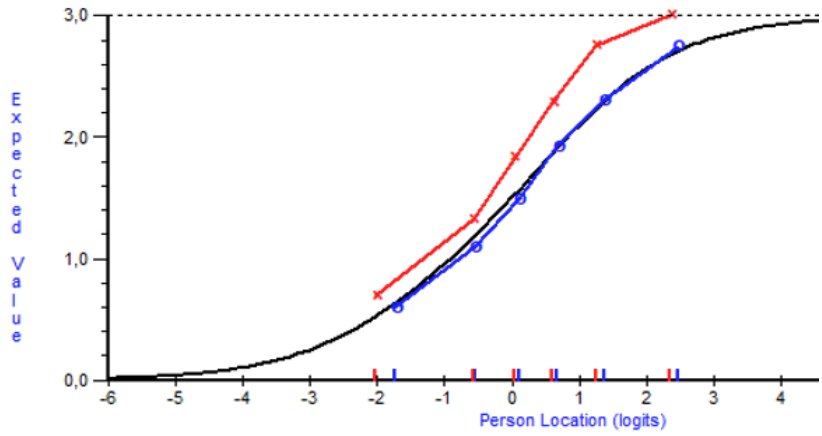


FIGURE 5

Graphical display of analysis of DIF, a situation when SvA students consistently score higher than Sv students on the given item (SV represented by blue line, SvA red line).

Thus, the item works differently among students following the syllabus for Swedish as a second language (SvA, represented by red line) compared to those following the syllabus for Swedish (Sv, represented by blue line). Students following the SvA syllabus consistently score higher than students following the Sv syllabus at the same location (ability) across the entire latent trait. Additionally, the comparison differs along the latent trait; the differences increase with higher abilities. Thus, the differences are largest among high achievers and smallest among low achievers. The implication of this finding is that it requires higher ability (knowledge based on the result on the complete test) for students following the Sv syllabus compared to those following SvA in order to pass this particular item. This particular item is a constructed response item. The DIF effect is highly significant ($F=18.49$; d.f.=1; $P<0.0001$), using ANOVA test of residuals. Thus, the null hypothesis (no differences between group means) needs to be rejected.

Figure 6 shows an example of an item where SvA students (red line) consistently score lower than Sv students (blue line). This is a selected response item. These differences span over the whole continuum but seem to decrease somewhat at the higher end of the latent trait, i.e., among high achievers. The difference is as large as one logit and thus significant ($F=44.74$; d.f.=1; $P<0.0001$), based on ANOVA analyses of residuals. Thus, the null hypothesis (no differences between group means) needs to be rejected.

In Figure 7, DIF has been resolved by syllabus group, i.e., the item was split into two different items (one for each syllabus group) in order to be able to quantify the differences between SvA students and Sv students. Thus, it is possible to reveal the actual size of the DIF effect, and to see how much higher ability is required for SvA students compared to Sv students in order to pass the item. In this case it is shown that for SvA students, more than 1 logit higher ability is required, compared to what is required for the Sv students.

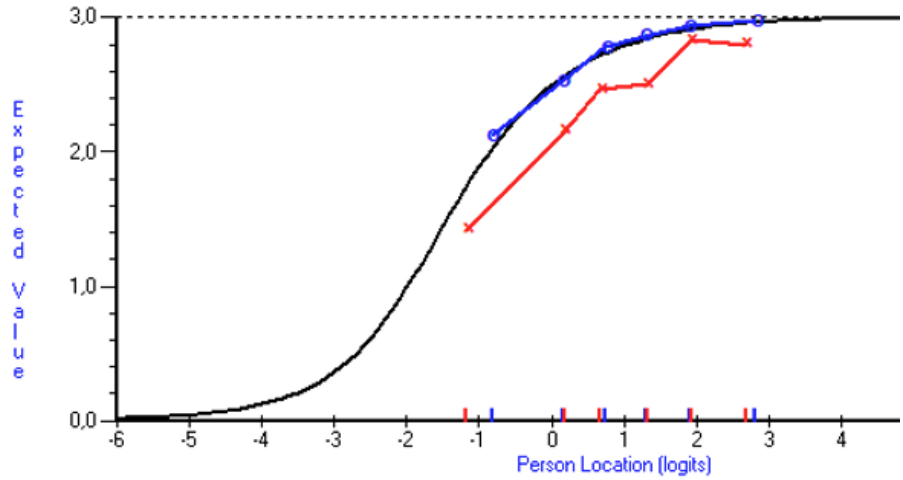


FIGURE 6

Graphical representation of DIF-effect where SvA students score consistently lower than Sv students (SV represented by blue line, SvA red line).

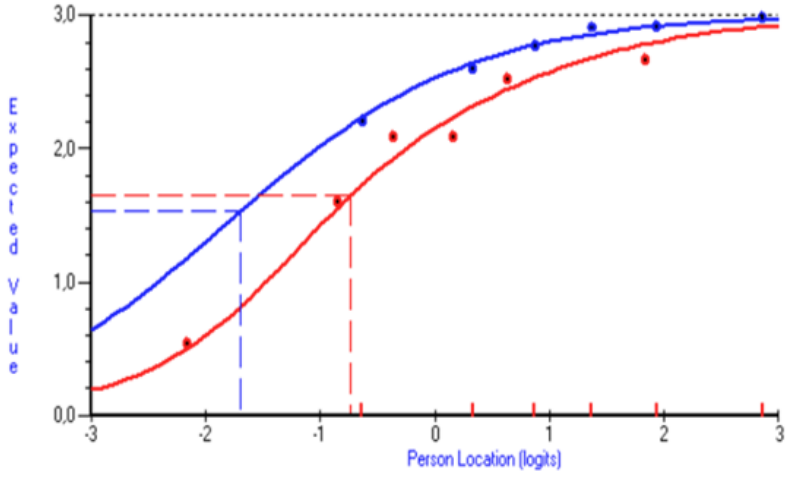


FIGURE 7

Same item as in figure 6 but now resolved for DIF by syllabus group in order to be able to quantify the DIF effect and to observe the actual size of the differences (SV represented by blue line, SvA red line).

Summary of results

It is important to recognize that "worse" or "better" in this context is *given* the same ability (knowledge based on the result on the complete test, as is understood within the framework usually applied within modern test theory).

As a group, students following the syllabus for SvA perform worse than students following the syllabus for Sv when looking at the overall results of the test, as is described earlier in Table 2.

However, analyzing DIF, when looking at different specific items and comparing the same ability level, there are systematic differences with different patterns derived.

In 18 percent of the tasks, SvA students performed worse than Sv students, i.e., almost every fifth task. However, in 10 percent of the tasks, SvA students performed better than Sv students.

TABLE 3

Tasks where SvA students perform significantly worse than Sv students

	Category	**Domain**	**Format**
1	Reference	Political system/Democracy	Selected response
2	Reference	Political system/Democracy	Selected response
3	Reference	General Social Studies	Selected response
4	Reference	Political system/Democracy	Selected response
5	Reference	Law	Selected response
6	Reference	Source criticism	Selected response
7	Reference	Media	Selected response
8	Reference	Political system/Democracy	Selected response
9	Reference	General Social Studies	Selected response
10	Reference	General Social Studies	Selected response
11	Reference	Media	Selected response
12	Analysis	Political system/Democracy	Selected response
13	Analysis	Political system/Democracy	Selected response
14	Analysis	Source criticism	Selected response
15	Analysis	Source criticism	Selected response
16	Analysis	Democracy	Constructed response
17	Analysis	Political system/Democracy	Constructed response
18	Analysis	Social issue	Constructed response

In Table 3 and Table 4, each row and number represent different items across the various tests. The columns indicate the type of task, labeled as reference or analysis, and domains, such as Economy, which allows us to categorize different types of task items. Previously, we hypothesized that domains would influence the DIF, but this was found to be insignificant. These specific labels were created for this study. Even though some items have the same labels (e.g. law), they can differ in other ways, for instance formats. The various items remain confidential, which is why we cannot specify them at this time.

According to the DIF analyses, most tasks where SvA students perform worse are those with a selected response format, as also summarized in Table 3.

The tasks in Table 3, taken from the four national tests, represent tasks where SvA students perform significantly worse: 61% are reference tasks, and 83% have a structured format, i.e., students respond in a closed (selected) response format, e.g., multiple-choice.

According to the DIF analyses, most tasks where SvA students perform better are those with an open response format (constructed response), which is also summarised in Table 4.

TABLE 4

Tasks where SvA students show significantly better results than Sv Students

	Category	Domain	Response format
19	Analysis	Economy	Constructed response
20	Analysis	EU	Constructed response
21	Analysis	Social issues/Media	Constructed response
22	Analysis	Economy	Constructed response
23	Analysis	Political system/Democracy	Constructed response
24	Analysis	Political system/Democracy	Constructed response
25	Analysis	Economy/Corruption	Constructed response
26	Analysis	EU	Constructed response
27	Reference	Law	Selected response
28	Reference	Economy	Selected response

80% of the items where SvA students perform significantly better than the Sv students are analysis tasks, and 80% are open in an open response format, i.e., where students write a short essay.

Discussion

The aim of this study was to analyze the impact that different formats can have on various subgroups of the Swedish national tests in Social Studies, using data from four years of investigation. A particular focus was directed towards the analysis of Differential Item Functioning (DIF) across the syllabus groups Swedish (Sv) and Swedish as a second language (SvA). These analyses were based on graphical comparisons of the item characteristics for Sv and SvA students but also complemented and confirmed by statistical tests using ANOVA of residuals between the groups.

On average, the performance of Sv students is higher compared to that of SvA students. However, the analysis conducted in this study, when analyzing the occurrence of DIF from a modern test theory approach, shows that there are indeed differences between the two syllabi groups when it comes to how the different task types included in the test work for the two groups. Thus, tasks using a selected response format (structured), asking for reference knowledge, are generally more difficult for the students following the syllabus of Swedish as a second language (SvA) compared to those following the general syllabus of Swedish, i.e. the majority of the students. The implication of this finding is that it requires much higher ability (knowledge based on

the result on the complete test, as is understood within the framework usually applied within the modern test theory approach) for SvA students to pass items of the selected response format than what is required for Sv students.

But there are items where SvA students perform better. These items are often in a constructed-response format and ask for analytical knowledge.

From a logical point of view, the findings of this study may be surprising, as one might expect that the group with the weakest language skills faces a disadvantage when responding to tasks that require them to formulate their own sentences in Swedish. Following that line of reasoning, the selected response items would then be easier for the SvA group, not more difficult. Therefore, it has previously been assumed that SvA students, due to poorer linguistic prerequisites, cope better with structured response formats. Thus, it has been hypothesized that closed (selected response) formats would provide better opportunities for these students to demonstrate their knowledge of the subject. However, this assumption is contradicted by the results of this study, which also is in line with findings reported from the first years of the Swedish national tests in geography (Alm Fjellborg & Molin, 2018). A hypothesized theoretical explanation of these results may be twofold: constructed response format (open-ended) gives the student more opportunity to explain the required knowledge in their own words, and the structured response formats tend to be linguistically very "compact" and force the student into a constraining specific terminology. Thus, if the student is not familiar with the terminology used in the task, it will be much more difficult to pass the task. These lines of reasoning are also concordance with the theoretical underpinnings of a Swedish study (Severin, 2002). Following Severin's (2002) reasoning, students would be better off using their own words, as is the case in the constructive response format of tasks. For these reasons there are empirical and logical explanations to why SvA students have more difficulties responding to selective response tasks, as the linguistic context is already set, and the student needs to be able to understand and apply the concepts given in the task.

The national tests in Sweden are now entering a new era as they are about to be digitally distributed and taken and are thus undergoing a major transformation. This is a consequence of the Swedish government commissioning the National Agency for Education to digitalize the national tests (Utbildningsdepartementet, 2017). The intention is also to increase the proportion of tasks using a selective response format. In the national test in Social Studies, about half of the tasks currently have a selective response format. Given the findings of this study, it may be possible to hypothesize that a substantial increase in the proportion of tasks using a selective response format would have consequences. Thus, students following the SvA syllabus, a group with lower achievements on the test, on average, may be particularly vulnerable given the proposed changes with increased proportions of selective response tasks. The findings from this study may be even more important in light of the intention to increase the importance of national tests in grading (SKOLFS 2022).

However, it is important to evaluate how different formats and contents interact in order to understand the links and consequences for specific groups. For example, Wilson (2023) emphasizes the importance of understanding the significance of formats

for exam accessibility. In addition, the results of this study may potentially have didactical implications relevant for class teachers as well as students' learning.

From a test developer perspective, this study stresses the importance of careful evaluations of Differential Item Functioning (DIF) and the interactions between different formats and contents for different groups of students, already in the development process (see Wilson, 2023, for a note on this matter). However, and this is a crucial point, the structure and tradition of each subject differ, the use of CR and SR items varies, and the subjects themselves differ. Given the consistent results of the study, it is possible to use the study findings in in the development process in order to facilitate fair assessment, and the results may also be generalizable, not only to other subjects, but it may also be relevant for other subgroups of students.

Ethical considerations

The data on which this study is based contains no personal information and therefore poses no threat to the personal privacy or integrity of the participants. Consequently, the reports and conclusions from the analysis is in the form that implies that no individuals can possibly be at risk of harm.

Methodological reflections

Given that the categorization of students into the SvA syllabus group is carried out by principals—and that the process is not fully transparent—there may be consequences for the study's results. For instance, some students may be deemed (by the principals) to have language skills too weak to be able to complete the test and are therefore excluded. Since the proportion of students following the SvA syllabus in the overall population is slightly higher than in this analyzed sample of test takers, this may indicate such exclusion. This, in turn, poses a risk of underestimation, if any, of the study's findings.

Funding

This study was conducted as part of the authors' positions as Assistant Project Manager and Scientific leader, respectively, for the national tests in Social Studies, at the University of Gothenburg.

Conflicts of interests

The authors have no conflict of interest, related to this study.

Conclusions and further research

In Sweden, the government has instructed the National Agency for Education to digitalize the national tests, but also to strive for a large proportion of the tasks to be automatically assessed, which requires a selected response format. Today, slightly less than half of the tasks have a constructed response (open-ended) format, so this change may have consequences. Changes in item format can have important implications for different groups of learners in the testing context. Changes involving more structured tasks may make an already vulnerable group of students (SvA) increasingly vulnerable to low achievements, which may be critical for their future opportunities. It is important to explore the interaction between task format and content in more depth - to find out why tasks with a particular format are more difficult for some groups than others. These insights should also have implications for teaching.

The national tests in Social Studies are high-stakes assessments. Teachers are already obliged to 'take special account' of the results of the national tests when grading the subject. Under the current proposal for a new grading system in Sweden, even greater emphasis will be placed on the national tests, and as the results it will account for 30% of the grades, and the average national test results at school level will be adjusted in relation to the school's grade point average. Given that the national test results are heavily influencing the final grades, the test results are also indirectly influencing student opportunities for Upper Secondary school admission

Even though it is not specifically addressed in this study, there is a lot of evidence suggesting that different types of tasks (and contents), in fact, may measure different dimensions, which we intend to elaborate on in future studies.

References

Alm Fjellborg, A., & Molin, L. (2018). Vilka typer av uppgifter gynnar elever som följer kursplanen i svenska som andraspråk? En undersökning med data från de nationella proven i geografi. *Acta Didactica* Norge, 12(4), Art. 5. https://doi.org/10.5617/adno.6286

Andersson, K., & Larsson, K. . (2023). Elevers perspektivbyten i svar på uppgifter från det nationella provet i samhällskunskap. *Nordidactica. Journal of Humanities and Social Science Education*, 13(2023:3), 66–84).

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561–573. https://psycnet.apa.org/doi/10.1007/BF02293814

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Educational Research and Perspectives*, 1982(9:1), 95–104.

Andrich, D. (1988). Rasch models for measurement. Sage. https://doi.org/10.4135/9781412985598

Andrich, D. (2011). Rating Scales and Rasch measurement. *Exp.Rev. Pharmacoeconomics Outcomes Res* 11(5), 571-585.

Andrich, D., & Marais, I. (2019). *A Course in Rasch Measurement Theory – Measuring in the Educational, Social and Health Sciences*. Singapore: Springer. https://doi.org/10.1007/978-981-13-7496-8

Andrich, D., Sheridan, B., & Luo, G. (2013). RUMM2030: *A windows program for the Rasch Unidimensional Measurement Model*. [Computer software] Perth, Western Australia: RUMM Laboratory.

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability*. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, Massachusetts: Addison-Wesley.

Børhaug, K, (2023). Kommentarer til Samfundsfag i Danmark fra et norsk perspektiv. *Nordidactica – Journal of Humanities and Social Science Education* 2023:1 60-65.

Christensen, AS, (2015). Demokrati-og medborgerskabsbegreber i grundskolens samfundsfag i Danmark, Norge, Sverige og Tyskland. *Nordidactica: Journal of Humanities and Social Science Education* 2015:1 64-92.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrica*, (16), 297–334. https://doi.org/10.1007/BF02310555

Gullberg, T (2023) Samfundsfag i Danmark –en finländsk kommentar och komparation, *Nordidactica: Journal of Humanities and Social Science Education* 2023:1 54-99.

Eklund, Niklas; Larsson, Anna, (2009). Samhällskunskap och disciplinfrågan – om utbildning av lärare i samhällskunskap. *Utbildning och demokrati*. Uppsala universitet 2009, vol 18, nr 1 69-91.

Hirsh, Å, (2015). *Nationella prov i grundskolan - En studie av hur lärare och rektorer uppfattar och hanterar prov och provresultat*, Skolverket 2015.

Landahl, Joakim (2015). Skolämnen och moralisk fostran: En komparativ studie av samhällskunskap och livskunskap, *Nordic Journal of Educational History* Vol. 2, no. 2 (2015), pp. 27–47. https://doi.org/10.36368/njedh.v2i2.55

Larsson, A., Ledman, K. (2023). Samhällskunskapsämnets utveckling i Sverige: kommentarartikel til Torben Spanget Christensen, Samfundsfag i Danmark, *Nordidactica: Journal of Humanities and Social Science Education*, 13(1): 66-73.

Lundahl, C., (2017), Perspektiv på nationella prov, *Utbildning & Demokrati*, 2017, vol 26, nr 2, 5–20. https://doi.org/10.48059/uod.v26i2.1078

Löfstedt, A, (2018), Det första nationella provet i samhällskunskap - en studie i bedömarsamstämmighet, *Acta Didactica Norge*, 12(4), Art. 13. https://doi.org/10.5617/adno.6283

Meyer, J. P. (2014). *jMetrik* (Version 4.1.1) [Computer software].

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980) with foreword and afterword by Benjamin D. Wright. Chicago: The University of Chicago Press.

Severin, R. (2002). *Dom vet vad dom talar om. En intervjustudie om elevers uppfattningar om begreppen makt och samhällsförändring*. Doktorsavhandling, Göteborgs universitet, Gothenburg Studies in Educational Sciences.

Skolförordningen, SFS 2011:185, 5 kap 9, 14 §.

Skolförordningen, SKOLFS 2022:185 (ändring).

Skolverket (2018a). *Skolverkets allmänna råd: Betyg och betygssättning*, Elanders Sverige AB, Stockholm.

Skolverket (2005), *Nationella utvärderingen av grundskolan*. Rapport 251.

Skolverket. (2011). *Läroplan för grundskolan, förskoleklassen och fritidshemmet 2011*.

Skolöverstyrelsen, K. (1962). *Läroplan för grundskolan 62*. Stockholm: Emil Kihlströms tryckeri.

Skolöverstyrelsen. (1969). *Lgr 69 Läroplan för grundskolan*. Stockholm: Bröderna Lagerström AB.

Skolöverstyrelsen. (1978). *Tillägg till Lgr 69*. Stockholm: Liber tryck.

Skolöverstyrelsen. (1980). *Lgr 80 läroplan för grundskolan*. Stockholm: Axlings Tryckeri AB.

Sandberg, E, (2021). *Att särskilt beakta nationella prov – en kvalitativ studie om hur lärare uppfattar och tillämpar förordningen om att särskilt beakta provresultatet vid betygssättning*, Uppsala universitet, Institutionen för pedagogik, didaktik och utbildningsstudier, 2021.

Spanget Christensen, T, (2023). Samfundsfag i Danmark – et overblik. *Nordidactica – Journal of Humanities and Social Science Education* 2023:1 35-53.
Utbildningsdepartementet. (1994). *Läroplaner för det obligatoriska skolväsendet och de frivilliga skolformerna.* Stockholm 1994.

Utbildningsdepartementet (2017), *Skrivelse U2017/03739/GV*. Uppdrag att digitalisera de nationella proven m.m. Stockholm 2017.

Wilson, M. (2023). *Constructing measures: An item response modeling approach* (2nd ed.). Routledge.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wu, M., Tam, P. H., & Jen, T-H. (2016). *Educational Measurement for Applied Researchers – Theory into Practice.* Singapore: Springer.

Digital sources

Eberly center for Excellence & Educational Innovation 2024
https://www.cmu.edu/teaching/assessment/basics/formative-summative.html
(accessed 2nd of September 2024).

Lärarnas historia, 2024
https://lararnashistoria.se/ (accessed 30th of October 2024).

Nationellt för centrum för svenska som andraspråk, 2024
https://www.su.se/nationellt-centrum-for-svenska-som-andrasprak/råd-och-stöd/vanliga-frågor?open-collapse-boxes=ccbd-vilkaeleverfårläsasva (accessed 28th of August 2024).