International Research in Geographical and Environmental Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rgee20

Sustainable development: Exploring gender differences in the Swedish national test in geography for grade 9

Andreas Alm Fjellborg & Kajsa Kramming

To cite this article: Andreas Alm Fjellborg & Kajsa Kramming (2022) Sustainable development: Exploring gender differences in the Swedish national test in geography for grade 9, International Research in Geographical and Environmental Education, 31:3, 172-187, DOI: 10.1080/10382046.2021.1927366

To link to this article: https://doi.org/10.1080/10382046.2021.1927366

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 18 May 2021.

Submit your article to this journal

Article views: 1377

View related articles

View Crossmark data

Citing articles: 2 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=rgee20

INTERNATIONAL RESEARCH IN GEOGRAPHICAL AND ENVIRONMENTAL EDUCATION
2022, VOL. 31, NO. 3, 172–187
https://doi.org/10.1080/10382046.2021.1927366

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Sustainable development: Exploring gender differences in the Swedish national test in geography for grade 9

Andreas Alm Fjellborg and Kajsa Kramming

Department of Education, Uppsala University, Uppsala, Sweden

ABSTRACT

This paper provides an analysis of how Swedish 15-year-olds perform on the high-stakes national assessments in geography. It explicitly addresses which item characteristics produce differential item functioning (DIF) in favor of boys and girls respectively. The findings show that DIF occurs in favor of girls in items with constructed response and primarily with content on the social dimension of sustainable development (SD), while boys are more favored by content outside the field of SD. The conclusions drawn are that content that reaches higher levels of Bloom's taxonomy favors girls, especially when the subject content concerns SD. This is important when analyzing the teaching and examination of sustainability issues in school.

KEYWORDS

DIF; education; geography; large-scale assessment; national tests; sustainable development; Sweden

Introduction

Responding to Lane and Bourke (2019, p.33) call, this research discuss what types of assessment instruments could *provide valid and reliable measures* of geographical knowledge. The aim of this paper is to reveal whether or not gender differences in school (Francis & Skelton, 2005; Voyer & Voyer, 2014) are mirrored in the large-scale standardized national tests in geography in Sweden. We focus on items related to sustainable development (SD), which is a focal point in the Swedish curriculum for geography as well as for Swedish education as a whole. Furthermore, SD items can be characterized as measuring high-level cognitive thinking. We test the extent to which types of content at different cognitive levels co-vary with the performance of boys and girls. In the large-scale standardized national tests in geography for 15-year-olds (grade 9 in compulsory school), various aspects of SD are tested and thus offer a way to analyze how Swedish 15-year-olds perform on such tasks in a high-stakes test context. This is important for developing our knowledge on how SD is studied and examined in the Swedish school system, and with what results. There are several research accounts that discuss SD in school (e.g., Olsson & Gericke, 2017). However,

CONTACT Andreas Alm Fjellborg andreas.alm.fjellborg@edu.uu.se Department of Education, Uppsala University, Uppsala 750 02, Sweden

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

these rarely make use of large-scale standardized test scores for understanding gender differences, performance levels or high-stakes examination results. We contribute to this literature by expanding the analysis of how content and examination form affect student performance levels.

Through differential item functioning (DIF) analysis, we map out gender differences in performance across all items and all years (2014-2019) for the national tests in geography in Sweden. In multivariate regressions, we test which item characteristics co-vary with the occurrence of DIF. We pose an overarching research question: Which item characteristics affect the occurrence of DIF favoring girls and boys respectively?

The paper starts by introducing sustainable development, and then presents a literature review of previous research about gender differences related to standardized tests and gender gaps in SD in education. This is followed by a summary of the Swedish national tests focusing on the subject of geography. The analytical approach is described in the methods section, followed by the empirical analysis. Finally, we discuss the findings.

Placing sustainable development in geography education

SD can be seen as a process of creating a world with global social equality without deteriorating the natural systems that are the basis for human existence (Kramming, 2017). The concept of education for sustainable development (ESD) is a way of incorporating SD in the school curriculum (United Nations [UN], 2002; United Nations Educational, Scientific & Cultural Organisation [UNESCO], 2014). SD can be taught in different ways serving different purposes. ESD strives to be a multi-faceted form of education that will train new generations of citizens and professionals in addressing sustainability issues (Kramming, 2017). One way of handling sustainability issues in education is through systems thinking, which means connecting elements from different disciplines and spatial scales to create an overall picture of sustainability challenges and possibilities (Wiek, Withycombe, & Redman, 2011). Systems thinking is especially important when analyzing relations between nature and humanity (Lezak & Thibodeau, 2016). In geography, such connections are central, making systems thinking a vital part of geographical knowledge and an essential part of the geography curriculum. Sustainable development focuses on problems that are *wicked* (Kramming, 2017; Kronlid, 2017). It means that the solutions to sustainability challenges are rarely conclusive, nor is there a *one solution fits all* approach that is applicable across geographical scales (Kramming, 2017; Rittel & Webber, 1973). This increases the need for students to engage in abstraction and discuss various pros and cons about certain aspects from a variety of perspectives.

Sustainability issues and systems thinking have implications for assessments in geography, at both local and standardized levels, since certain teaching and examination methods are required to impart and assess relevant knowledge and abilities.

Geography teaching follows three selective traditions (for Sweden, see Molin, 2006). The *first* is the fact-oriented tradition, which gives basic geographical

knowledge building on scientific facts. The *second* is the normative tradition focusing on students' environmentally friendly values and behaviors. The *third* is the pluralistic tradition, where students are trained in creative and critical thinking towards sustainability issues (Sandell, Öhman, & Östman, 2003; Sund, 2008). A study by Borg, Gericke, Höglund, and Bergman (2012) shows that a normative perspective on SD is common among Swedish teachers. The normative tradition is based on teaching students environmentally friendly behaviors and values and explaining why it is important to consider people living in other parts of the world in less privileged circumstances than most people do in Sweden. Teaching that follows this tradition may evolve around the environmental and social dimensions of sustainable development lifting vulnerable and marginalized groups in society such as women, girls and children. Notably, the economic dimension is often neglected, which results in students and teachers being more familiar with the social and environmental dimensions of SD (Olsson & Gericke, 2016; Pettersson, 2014). The pluralistic tradition is the least represented when geography is taught in Swedish schools, indicating that ESD is not yet fully in use (Molin, 2006; Torbjörnsson & Molin, 2014).

Taking all the selective traditions into consideration, teaching and learning SD entail different levels of cognitive thinking, ranging from recalling and explaining facts to synthesizing knowledge from different areas. This development follows Bloom's knowledge taxonomy (Anderson, Krathwohl, & Airasian, 2001; Bloom, 1956).

According to Bloom's taxonomy (Anderson et al., 2001; Bloom, 1956), levels of cognition start with the abilities to remember, understand and apply knowledge, whereupon the student analyzes, evaluates and finally creates new knowledge by synthesizing different elements (Anderson et al., 2001).

The different cognitive levels can be assessed with different item formats, where the first steps of the taxonomy are more suitable to test with selected-response items, such as multiple-choice or matching (MC items). Higher levels of the taxonomy are more suited for open-ended response items such as constructed response (CR items) (Haladyna, 2004; Haladyna, Downing, & Rodriguez, 2002; Rodriguez, 2016; Wikström, 2013). However, some argue that higher-level cognitive skills can also be assessed by selected response (Scully, 2017). Bijsterbosch, van der Schee, and Kuiper (2017) show that summative geography tests in the Netherlands often test only lower-level cognitive demands (for the U.S., see Edelson, Shavelson, & Wertheim, 2013). Higher-order cognitive demands are more often found in formative tests (Bijsterbosch et al., 2017). This raises questions about how items regarding SD work in summative tests where the content is framed in line with the pluralistic tradition, thus encouraging higher-level thinking.

School and assessment performance – explaining gender gaps

Overall, girls perform better at school than boys do (Francis & Skelton, 2005; Skolverket [Swedish National Agency for Education], 2018; Voyer & Voyer, 2014). Results from the PISA tests show that girls tend to perform better in reading, while boys perform better in math (Hermann & Kopasz, 2019). However, in Sweden, girls

outperform boys in reading but score on an equal level in math (Skolverket [Swedish National Agency for Education], 2018). Previous research offers four factors that affect the gender gap in assessment results in reading and math among 15-year-old students. The *first* is how the educational system is organized in terms of selection into educational tracks. Later selection benefits girls and creates a larger gender gap (van Hek, Buchmann, & Kraaykamp, 2019). The *second* factor is whether teaching is tailored to individual students and is less standardized. In those cases, all students show higher levels of reading skills, but the gender gap in favor of girls is wider (Hermann & Kopasz, 2019; van Hek et al., 2019). *Third*, in countries where government regulations determine curricula, reading scores are even more in favor of girls (Ayalon & Livneh, 2013; van Hek et al., 2019). *Fourth*, in countries that offer standardized testing, such as national tests in different subjects, the gap between girls and boys in math is narrowed (Ayalon & Livneh, 2013). In sum, this means that girls have relative benefits in terms of succeeding in school in educational systems with late selection, individualized teaching methods, governmentally supported curricula and standardized tests. Sweden is a country that fills all these criteria.

In some subjects (e.g., mathematics and science), the gender gap is not present, and some findings even show that boys perform better (Liu & Wilson, 2009; Voyer & Voyer, 2014). This could be due to variations between genders with regard to anxiety, stereotyping (Herts & Levine, 2020), spatial thinking (Wai, Lubinski, & Benbow, 2009) or motivation (Butt, Weeden, Chubb, & Srokosz, 2006). Since most internal school-based examinations test lower cognitive levels, such as remembering facts and concepts (Bijsterbosch et al., 2017), students might be ill prepared for summative examinations where *evaluation* or *creating* on higher cognitive levels are requested.

In geography, it is likely that girls outperform boys (as in environmental education, e.g., Boeve-de Pauw, Jacobs, & Van Petegem, 2014). However, boys perform better than girls when it comes to certain types of knowledge, for example, place names and locations (although gaps are narrowing in Sweden, see Hennerdal, 2016), which is in line with findings applicable to spatial thinking skills (Tomaszewski, Vodacek, Parody, & Holt, 2015). Butt, Weeden, and Wood (2004) show that boys do not perform as well as girls on written exams in geography. According to them, assessment methods are important for understanding gender gaps, which might be related to preferred styles of learning in the classroom. Likewise, it may be related to the preferred style of teaching, regarding different types of geography content. In Bednarz and Lee (2019) research review, there are no instances where girls outperform boys on spatial abilities, but a sizeable number of studies show no gender difference.

The gender gap appears to be especially significant when examining MC and CR items (Lane, Wang, & Magone, 2005; Lyons-Thomas, Sandilands, & Ercikan, 2014). The item format is useful also when analyzing performance in literacy (Schulz-Heidorf & Støle, 2018). Girls outperform boys on CR items, and boys do better on MC items. This has led researchers to argue for mixed-format tests to avert gender bias (Kacprzyk, Parsons, Maguire, & Stewart, 2019). When testing only MC items, Stiller et al. (2016, p. 723) find that the length of response options, visual images, formulas, abstract and specialist concepts increase item difficulty, and that the inclusion of tables and longer item stems reduce item difficulty.

Furthermore, using the PISA assessment, Le Hebel, Montpied, Tiberghien, and Fontanieu (2017) show that the cognitive demands influence item difficulty, while there are inconclusive results on whether or not different stimuli are influential. CR items seem more difficult while MC items tend to reduce item difficulty (see also Mullis, Martin, & Foy, 2013).

The Swedish national tests

In Sweden, school is compulsory for ten years, with grade 9 being the highest level. National tests are given in grades 3, 6 and 9 and impact students' final grades (in 6th and 9th grades) since test results should be given considerable weight when teachers grade their students (Sveriges Regering [Government of Sweden], 2017). After compulsory school, students can apply to upper secondary school, where different programs have different entry levels. There are both vocational and theoretical programs that prepare students for higher academic studies at university. The national test results do affect the grades and thus the possibilities for choosing an upper secondary school program.

Since the 2012-2013 school year, national tests in the social science subjects civics, history, religion and geography have been given in the final year of Sweden's compulsory school (with 15-year-old students); earlier, only English, math and Swedish were tested. The cohort of students taking the tests in social sciences are divided into four groups, and each group is given one of the social science tests.

The starting point for the development of national tests is the curriculum for compulsory school and the curriculum in each subject. The assessments cover as much of the subject curriculum as possible. Further, all tests must be reliable and display high inter-rater reliability, which is secured by the development of rating schemes.

The Swedish National Agency for Education (Skolverket) stipulates that the aim of the tests is to support fair and equivalent grading of students across schools. The grading process in which individual teachers assess their own students, or where teachers work in groups to assess students within their school, elevates the risk of inconsistencies in grading between schools and teachers (Gustafsson & Erickson, 2018).

Each item in the geography test is rated on a scale that begins at F (not acceptable) followed by E (pass), and then C and A (pass with distinction). The E, C and A marks on items are understood as evidence of a student's competence in relation to the knowledge demands for each grade in the curriculum.[1] The test contains a mix of MC items (stacked into items graded F, E, C or A) and CR items. For the geography subject assessment, content is guided by the four abilities (Table 1) that geography education is intended to develop according to the curriculum (Skolverket [Swedish National Agency for Education], 2011).

In the tests, the different abilities are represented to different extents based on the knowledge demands in the curriculum. The variation and layout of the tests are important factors; as Thornes (2004) and Mukherjee (2015) show, students perform better if items are supported by illustrations or other forms of supportive arrangements (e.g., concepts, perspectives to be used). There are often stimuli and resources

Table 1. The four abilities in the geography curriculum.

The student should develop the ability to:	Ability name
… analyze how nature's own processes and human activities shape and reshape living environments in different parts of the world	Geographical processes
… explore and analyze the interplay between humans, society and nature in different parts of the world	Interplay human-nature
… make geographical analyses of the world around us and assess results with the aid of maps and other geographical sources, theories, methods and techniques	Geographical analysis
… assess solutions for various environmental and developmental issues from the basis of considerations of ethics and SD	Assess solutions

Source: Skolverket [Swedish National Agency for Education] (2011).

Table 2. Descriptive statistics of national assessment items in geography, 2014-2019.

	N	%
Students mean N	1636	
Total items N	172	
Subject ability		
Geographical processes	52	30%
Interplay human-nature	28	16%
Geographical analysis	63	37%
Assess solutions	29	17%
DIF		
DIF girls	27	16%
DIF boys	40	23%
SD		
SD total	70	41%
SD climate	10	6%
SD environment	19	11%
SD economic	10	6%
SD social	31	18%
Subject content		
Concepts	20	12%
Energy	10	6%
Field studies	10	6%
Trade	14	8%
Climate	36	21%
Environment	15	9%
Place names	16	9%
Social	51	30%
Item types		
MC items total	84	49%
Dual choice (MC)	27	16%
Multiple-choice (MC)	21	12%
Short written (MC)	5	3%
Match (MC)	31	18%
CR items total	88	51%
Explain (CR)	19	11%
Reasoning (CR)	69	40%
Stimuli/resources		
Concepts	30	17%
Pictures	91	53%
Graphs	39	23%
Maps	101	59%
Models	12	7%
Perspectives	33	19%

Source: Authors' calculations using the NP-GEO database, Uppsala University.

such as graphs, maps or pictures that may help students answer the items. Further, in many CR-items, students are asked to build their answers from a set of perspectives or concepts. For example, when asked about what consequences production of electronic waste might have, students are helped to frame their answer by using perspectives such as "Local-Global," "The rich–The poor" or "Today–In the future."

The Swedish national tests in geography are large-scale standardized tests with a summative function, but they include items across the whole scale of cognitive levels.

From the year 2022 onwards, both the curriculum and the scoring of the tests will change due to a new revised geography curriculum and digitalization of the tests. This study uses test results from the pen-and-paper exams given from 2014 to 2019.

Methods and data

We used the NP-GEO database at Uppsala University. This database contains a sample of students' results on all items in the national tests in geography for grade 9 from 2014 to 2019. The sample collected contains about 1500 complete test forms and information on gender and language courses in Swedish of the students taking the test for each year. The teachers register students born on two dates each month in a web-based form.

Using a differential item functioning (DIF) analysis, we identified systematic differences between boys and girls that could not be explained by the general level of knowledge differences between these groups. The analysis shows whether two student population groups have varying results irrespective of individuals' ability to answer a test item correctly. When a test item measures the same ability across population groups, results should only differ according to the individual test takers' abilities (Tennant & Pallant, 2007). DIF does not mean that the item necessarily has item bias that makes one group perform better than expected, but it shows that the item does not work as anticipated (Martinková et al., 2017; Zieky, 2003). Items displaying DIF should always be scrutinized, but not necessarily discarded.

The DIF analysis compares a reference group, normally the majority population group, with a minority group. In this paper, girls are compared with boys. There are several different ways to test items for DIF. In this paper, the Mantel-Haenszel (MH) technique (χ 2) is used. Other relevant ways are Item Response Theory (IRT) measures and logistic/ordinal regressions (e.g., Crane, Gibbons, Jolley, & Van Belle, 2006; Zumbo, 1999).

Table 2 displays descriptive statistics of all items in the geography tests from 2014 to 2019. We do not distinguish between uniform and non-uniform DIF. Uniform DIF occurs when differences are across all competencies while non-uniform DIF is when DIF occurs only among students on parts of the competence scale.

After item DIF is analyzed, logistic regressions are used to analyze what types of item characteristics increase the probability that an item displays significant DIF in favor of girls or boys. In logistic regressions, odds ratios are estimated. These represent the association between the independent and dependent variables. Odds ratios above 1 indicate a positive association, while odds ratios below 1 indicate a negative association. These models serve to draw out the types of item characteristics that affect the occurrence of DIF in

Table 3. Results from DIF analysis. All 2014-2019 items are from national assessments in geography, grade 9, Sweden.

		Type of DIF		
		No DIF	DIF favors girls	DIF favors boys
Subject Ability	Geographical processes	38	2	12
	Human-nature interplay	23	5	0
	Geographical analysis	33	2	28
	Assess solutions	11	18	0
Format	CR	60	27	1
	MC	45	0	39
Sustainable Development	Yes	44	25	1
	No	61	2	39
Type of SD	Social	17	14	0
	Economic	5	5	0
	Environmental	15	4	0
	Climate	7	2	1

Source: Authors' calculations, NP-Geo database.

favor of girls and boys respectively. All independent variables (with definitions) are presented in Table 2. In the descriptive analysis prior to the regression models, we analyze item characteristics that earlier research found important for item difficulty: format (MC or CR), content and stimuli. Further, the dimensions of SD, environmental, social and economic sustainability, are included in the analysis. We distinguish between climate change and environmental sustainability items.

Five aspects from the descriptive statistics of items in the national tests in geography in Table 2 deserve to be highlighted. First, a minority of assessment items display DIF. Second, when it comes to the SD items, the tests emphasize the social and environmental aspects of SD; the economic aspect is not represented to the same extent. Third, the tests are more concerned with social issues compared to other types of subject content. Fourth, the representations of CR and MC items are rather similar, and fifth, the tests make extensive use of pictures, maps and graphs as resources or stimuli.

Results

Table 3 presents the outcome of the DIF analysis. Girls tend to perform better than boys on the tests. In fact, girls outperformed boys in all years of this analysis, with the largest difference in 2017-2018. However, the number of items favoring boys is higher, with 40 items for boys and 27 for girls out of 172 items. Items favoring girls are mainly in the subject ability of assessing solutions to sustainability. Geographical analysis and geographical processes seem to produce DIF in favor of boys. Interestingly, the item format clearly has an impact. All items producing DIF in favor of girls are CR items, while those items producing DIF in favor of boys are all MC items. Girls also seem to be most favored by items in the social dimension of SD.

Items relating to SD have become more prominently featured in the national tests in geography over time. In 2014, eight items (28%) were categorized as SD items, and the number peaked in 2018 (14 items (47%)).

Table 4 shows the differences in performance between boys and girls. Girls tend to perform better than boys do on all SD items. There are a few areas in geography

Table 4. Mean total points achieved on items, by category 2014-2019.

	Mean (SD) N Girls	Boys
Non-ESD	1.32(0.45)102	1.32(0.45)102
SD total	1.60(0.20)70	1.36(0.19)70
SD social	1.58(0.21)31	1.32(0.20)31
SD economic	1.56(0.14)10	1.31(0.16)10
SD environmental	1.58(0.24)19	1.38(0.18)19
SD climate	1.72(0.09)10	1.52(0.13)10
Concepts	1.37(0.53)20	1.42(0.53)20
Energy	1.39(0.34)10	1.31(0.27)10
Field studies	1.44(0.39)10	1.36(0.36)10
Trade	1.44(0.27)14	1.24(0.20)14
Climate	1.37(0.45)36	1.32(0.43)36
Environment	1.54(0.31)15	1.32(0.22)15
Place names	1.57(0.34)16	1.63(0.37)16
Social	1.44(0.36)51	1.27(0.30)51

Notes: Total items 172, 70 SD items. The scoring scale has been altered from the F, E, C, A scale to a 0-3 point scale.
Source: Authors' calculations, NP-Geo database.

where boys outperform girls; place names, locations and geographical concepts display higher mean score values for boys compared to girls. However, it should be noted that both boys and girls fare rather well on SD items, even if the difference between the genders is wider in SD items compared to other items.

To sum up, girls tend to outperform boys in most of the subject categories and especially in aspects of SD. Girls fare better on CR items, where students are required to provide a written answer, while boys perform better on MC items, especially those concerning place names. As observed earlier, aspects relating to the first (lower) cognitive levels – remembering, understanding and applying – favor boys, at least in part, whereas girls tend to perform better in the higher cognitive levels of analysis, evaluation and synthesizing. We continue to use the results from the DIF analysis to determine whether these tentative conclusions hold even when controlling for a range of item characteristics that could affect the occurrence of DIF.

Explaining gendered DIF in the national assessments in geography

To be able to explore what item characteristics produce DIF among items in the national tests in geography, we ran a set of logistic regression models. The dependent variable in this first set of models is a binary variable separating the items that display DIF in favor of girls from items that do not. Results are shown in Table 5.

The binary variable indicating whether or not an item is related to SD shows that SD items are more likely to display DIF in favor of girls compared to items that do not specifically relate to SD (Table 5, model 1). This corroborates the findings in the descriptive analysis. In the second model, we use a categorical SD-type variable instead of the binary variable used in model 1. The results indicate that social aspects have an impact on the likelihood of an item indicating DIF in favor of girls. Results are significant even when controlling for a range of other factors as item difficulty and stimuli/resources. Since there is no SD item that has a MC format, nor any MC items displaying DIF in favor of girls, it is not possible to analyze the impact of item format in this regard.

Table 5. Odds ratios from logistic regression. Dependent variable: Item displays DIF in favor of girls. All items (N = 172) from national assessments in geography, Sweden, grade 9, 2014-2019.

Variable type	Variable	Model 1 Exp(B)(S.E.)Sig.	Model 2 Exp(B)(S.E.)Sig.
	Year	Yes	Yes
Item controls	Item number	1.006(0.071)	1.047(0.076)
	Share (%) fail	0.981(0.043)	0.951(0.048)
Stimuli/resources	Pictures	1.181(0.586)	1.31(0.611)
	Maps	1.226(0.550)	0.84(0.623)
	Graphs	1.665(0.696)	1.174(0.744)
	Reasoning	8.341(1.002)*	12.793(1.082)*
SD	About SD	7.603(1.024)*	
SD dimension	Climate		2.727(1.333)
	Environment		1.551(1.294)
	Economic		6.579(1.274)
	Social		9.535(1.04)*
	Constant	0.015(1.406)**	0.02(1.437)**
Nagelkerke R Square		0.431	0.476
−2 Log likelihood		99.917	93.849

Notes: Sig. $p > 0.001$ = ***, $p > 0.01$ = **, $p > 0.05$ = *
Source: Authors' calculations, NP-Geo database.

Table 6. Odds ratios from logistic regression. Dependent variable: Item displays DIF in favor of boys. All items (N = 172) from national assessments in geography, Sweden, grade 9, 2014-2019.

Variable type	Variable	Model 1 Exp(B)(S.E.)Sig.	Model 2 Exp(B)(S.E.)Sig.
	Year	Yes	Yes()
Item controls	Item number	1.006(0.057)	1.009(0.059)
	Share (%) fail	0.99(0.024)	1.014(0.028)
Stimuli/resources	Pictures	1.573(0.528)	1.412(0.545)
	Maps	2.488(0.731)	2.14(0.758)
	Graphs	1.974(0.661)	1.228(0.736)
Content type	About SD	0.055(1.109)**	0.383(1.43)
	Social (ref.)		
	Concepts	11.493(0.973)*	5.489(1.029)
	Energy	5.258(1.260)	3.366(1.331)
	Field study	0.646(1.110)	1.022(1.156)
	Trade	2.523(1.389)	4.159(1.504)
	Climate	3.607(0.829)	3.222(0.856)
	Environment	0.000(9243.16)	0(9255.439)
	Place names	76.856(1.089)***	48.538(1.095)***
	MC (CR = ref.)		19.821(1.339)*
	Constant	0.016(1.253)**	0.002(1.769)***
Nagelkerke R Square		0.553	0.596
−2 Log likelihood		108.042	100.089

Notes: Sig. ***= $p > 0.001$, **= $p > 0.01$, *= $p > 0.05$.
Source: Authors' calculations, NP-Geo database.

Girls perform better on SD items, exceeding what may be anticipated given overall test results. The social realm of SD seems to be particularly important.

For a complete analysis of DIF and gender, a similar logistic regression model was fitted to the data with a binary dependent variable indicating whether the item displays DIF in favor of boys (Table 6). It is important to remember that the descriptive statistics indicated that place names and concepts were correlated with higher results for boys compared to girls, and DIF in favor of boys was primarily found in MC items.

SD items have a negative impact on the occurrence of DIF in favor of boys. Instead, boys perform better than expected on MC items in general, particularly items

Table 7. OLS regression. Dependent variable difference in boys and girls item score. All items (N = 172) from national assessments in geography, Sweden, grade 9, 2014-2019.

Variable type	Variable	B(S.E.)Sig.
	Year	Yes
Item controls	Item number	−0.003(0.003)
	Share (%) fail	0.001(0.001)
Stimuli/resources	Pictures	0.022(0.022)
	Maps	0.03(0.03)
	Graphs	0.039(0.025)
Content type	About SD (yes = 1)	−0.068(0.035)
	Social (ref.)	
	Concepts	0.062(0.044)
	Energy	0.058(0.048)
	Field study	−0.087(0.052)
	Trade	0.051(0.042)
	Climate	0.062(0.031)*
	Environment	−0.004(0.041)
	Place names	0.099(0.044)*
	MC (CR = ref.)	0.094(0.035)**
	Constant	−0.251(0.054)***
R2		0.512

Notes: Sig. ***= $p > 0.001$, **= $p > 0.01$, *= $p > 0.05$.
Source: Authors' calculations, NP-Geo database.

concerning place names (Table 6). Both these aspects are related to testing knowledge in the lower levels of Bloom's taxonomy and do not relate to SD.

Robustness check

The logistic regression models above are not the only way to identify the variables that influence differences between boys' and girls' assessment performance. As an additional modelling strategy, we employ a linear regression specified with a dependent variable that represents the difference of the mean scores of boys and girls on each test item. Negative values of the coefficients in Table 7 show that girls tend to perform better, while positive values indicate that boys perform better. The results confirm those from the logistic models using the binary DIF variables. Boys perform better on MC items and content regarding place names. That includes MC items about climate, which are rather difficult and fall under the cognitive level of analysis. On those items boys perform better, but there was no source of DIF originating from this item content. The variable indicating SD content does not have a significant impact on item difference in item performance between boys and girls. The coefficient indicates the direction of correlation that we expect from the analysis above. This finding indicates that boys and girls perform very well on SD items, as indicated by the descriptive findings above. However, the DIF analysis above shows that girls perform better than expected.

Conclusions

This study shows that girls perform better than expected when asked to reason about a problem and on test items concerning SD, particularly regarding the social domain of SD. The results are in line with results from the PISA tests showing higher literacy

skills among girls (Hermann & Kopasz, 2019), and previous studies on environmental education (Boeve-de Pauw et al., 2014). We show that this pertains to the social domain of SD. These are all abilities belonging to the higher levels in Bloom's taxonomy. SD items have a negative effect on the occurrence of DIF in favor of boys. However, the difference between boys and girls in item scores on SD items is smaller compared to other item content. The findings raise some discussion points.

It is clear that both item content and item format affect how boys and girls perform on different test items (content, e.g., Boeve-de Pauw et al., 2014; format, e.g., Kacprzyk et al., 2019; Schulz-Heidorf & Støle, 2018). Our study shows an independent effect of SD content in favor of girls. Sustainability issues and assessment of SD require abstraction and engagement with higher-order cognitive skills such as analysis, evaluation and synthesizing. Girls seem to do much better in these areas compared to boys. An inclusive assessment of SD issues and 15-year-olds' skills should consider the possible causes of these findings. We identify three plausible causes. First, previous studies show that girls outperform boys in reading, which might lead girls to make better sense of what is actually being asked (e.g., question and instructions). Second, since girls outperform boys in reading, they may also outperform them in writing, which would have a clear impact on their results on SD items. Investigations into how much time girls and boys spend on each test item respectively have yet to be done, but such studies could shed light on whether girls linger on SD items, allowing them to elaborate their answers. Finally, the aspect of the social realm of SD being the most normative in the classroom (Borg et al., 2012) makes it the most "open-ended" item type, thus girls may once again be favored due to stronger writing and reasoning skills.

It is further likely that the social realm of SD is taught and examined in standardized testing in a way that maintains gendered socialization processes (Olsson & Gericke, 2017). It is plausible that social sustainability items favor girls due to stronger interest in topics about the development of girls' and women's education, reproductive rights and position in societies in different parts of the world. If this is an effect of the normative selective tradition being the strongest in the classroom, teaching about SD may have to focus more on the pluralistic selective tradition and systems thinking. A broader scope of SD, illuminating also environmental and economic dimensions, could result in more students (boys and girls) finding a larger selection of topics of interest when faced with assessment items about SD. In the national tests in geography in Sweden, all three dimensions of SD are represented. Since girls fare better on all SD items, current teaching traditions in Swedish schools should be more closely examined.

Girls score better on literacy tests compared to boys (Hermann & Kopasz, 2019), which might imply that they also outperform boys in writing. Hence, one limitation of this study is that we cannot know whether girls actually have better knowledge about SD or if they merely express that knowledge more thoroughly than boys do. This might affect the students insofar as the national tests in geography are to be given considerable weight in grading. One reason to choose to examine SD issues with CR formats is that higher cognitive levels are being tested. The CR format makes it possible to evaluate students'systems thinking abilities in a way that is unfeasible

with MC formats, since it is difficult to frame the wickedness of SD issues in MC questions and distractors suitable for 15-year old students. The difficulty lies in developing MC items that allow for examination of the higher levels of Bloom's taxonomy, granting students the possibility to show what they know at the same time as the items are construct relevant.

This paper was partly prompted by the lack of statistical knowledge about assessment results concerning SD in Swedish education (Statistics Sweden, 2017). We conclude that the mean score on SD items tends to be higher than on other types of items for both boys and girls, but that the achievement gap between genders is generally stronger in SD items compared to other content. Girls perform better than expected on these items, and this may be attributed to item content. SD has a negative effect on the likelihood of an item displaying DIF in favor of boys. Instead, boys performed better on items about place names and MC items.

We show that Swedish 15-year-old students tend to be very knowledgeable about SD. While it may not be achievable to develop MC items on SD, which probably would benefit boys, we propose that SD should be incorporated into teaching in a more pluralistic way. This relates to the argument from Butt et al. (2004) referring to how preferred styles of learning may affect examination results, and thus offering a broader spectrum of topics that might interest both boys and girls preparing them for high-stakes summative assessments.

Note

1. A grade of D is received if the student has demonstrated most of the C-level knowledge requirements but has not reached this threshold. Similarly, a grade of B is given to a student who does not quite reach the A level.

Disclosure statement

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

References

Anderson, L. V., Krathwohl, D. R., & Airasian, P. W. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.

Ayalon, H., & Livneh, I. (2013). Educational standardization and gender differences in mathematics achievement: A comparative study. *Social Science Research*, *42*(2), 432–445.

Bednarz, R., & Lee, J. (2019). What improves spatial thinking? Evidence from the spatial thinking abilities test. *International Research in Geographical and Environmental Education*, *28*(4), 262–280.

Bijsterbosch, E., van der Schee, J., & Kuiper, W. (2017). Meaningful learning and summative assessment in geography education: An analysis in secondary education in the Netherlands. *International Research in Geographical and Environmental Education*, *26*(1), 17–35.

Bloom, B. S. (1956). Taxonomy of education objectives. In *The classification of educational goals. Handbook I: Cognitive domain*. New York: McKay, *20*, 24.

Boeve-de Pauw, J., Jacobs, K., & Van Petegem, P. (2014). Gender differences in environmental values: An issue of measurement? *Environment and Behavior*, *46*(3), 373–397.

Borg, C., Gericke, N., Höglund, H.-O., & Bergman, E. (2012). The barriers encountered by teachers implementing education for sustainable development: Discipline bound differences and teaching traditions. *Research in Science & Technological Education*, *30*(2), 185–207.

Butt, G., Weeden, P., Chubb, S., & Srokosz, A. (2006). The state of geography education in English secondary schools: An insight into practice and performance in assessment. *International Research in Geographical and Environmental Education*, *15*(2), 134–148.

Butt, G., Weeden, P., & Wood, P. (2004). Boys' underachievement in geography: An issue of ability, attitude or assessment? *International Research in Geographical and Environmental Education*, *13*(4), 329–347.

Crane, P., Gibbons, L., Jolley, L., & Van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical Care*, *44*(Suppl 3), S115–S123. www.jstor.org/stable/41219511.

Edelson, D. C., Shavelson, R. J., & Wertheim, J. A. (Eds.). (2013). *A road map for 21st century geography education: Assessment*. Washington, DC: National Geographic Society.

Francis, B., & Skelton, C. (2005). *Reassessing gender and achievement: Questioning contemporary key debates*. New York, NY: Routledge.

Gustafsson, J. E., & Erickson, G. (2018). Nationella prov i Sverige–tradition, utmaning, förändring. *Acta Didactica Norge*, *12*(4), 2–20.

Haladyna, T. M. (2004). *Developing and validating multiple-choice items* (3rd ed.). New York, NY: Routledge.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for a classroom assessment. *Applied Measurement in Education*, *15*(3), 309–334.

Hennerdal, P. (2016). Changes in place location knowledge: A follow-up study in Arvika, Sweden, 1968 and 2013. *International Research in Geographical and Environmental Education*, *25*(4), 309–327.

Hermann, Z., & Kopasz, M. (2019). Educational policies and the gender gap in test scores: A cross-country analysis. *Research Papers in Education*, 1–22. https://doi.org/10.1080/02671522.2019.1678065.

Herts, J., & Levine, S. (2020). Gender and math development. In *Oxford research encyclopedia of education*, Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.1186.

Kacprzyk, J., Parsons, M., Maguire, P., & Stewart, G. (2019). Examining gender effects in different types of undergraduate science assessment. *Irish Educational Studies*, *38*(4), 467–480.

Kramming, K. (2017). Miljökollaps eller hållbar framtid? Hur gymnasieungdomar uttrycker sig om miljöfrågor, Geografica 13. In *Doktorsavhandling Kulturgeografiska Institutionen*. Uppsala: Uppsala universitet.

Kronlid, D. (2017). *Skolans värdegrund 2.0: Etik för en osäker tid*. Stockholm: Natur & kultur.

Lane, R., & Bourke, T. (2019). Assessment in geography education: A systematic review. *International Research in Geographical and Environmental Education*, *28*(1), 22–36.

Lane, S., Wang, N., & Magone, M. (2005). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and Practice*, *15*(4), 21–27.

Le Hebel, F., Montpied, P., Tiberghien, A., & Fontanieu, V. (2017). Sources of difficulty in assessment: Example of PISA science items. *International Journal of Science Education*, *39*(4), 468–487.

Lezak, S. B., & Thibodeau, P. H. (2016). Systems thinking and environmental concern. *Journal of Environmental Psychology*, *46*, 143–153.

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education*, *22*(2), 164–184.

Lyons-Thomas, J., Sandilands, D., & Ercikan, K. (2014). Gender differential item functioning in mathematics in four international jurisdictions. *Education & Science/Egitim ve Bilim*, *39*(172), 20–32.

Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, *16*(2), rm2.

Molin, L. (2006). *Rum, frirum och moral: En studie av skolgeografins innehållsval, Doktorsavhandling, Geografiska regionstudier. nr. 69*. Uppsala: Kulturgeografiska institutionen, Uppsala universitet.

Mukherjee, S. (2015). Towards visual geography: An overview. *Practicing Geography*, *19*(2), 13–22.

Mullis, I. V. S., Martin, M.O., & Foy, P. (2013). The impact of reading ability on TIMSS mathematics and science achievement at the fourth grade: An analysis by item reading demands. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade – Implications for early learning* (pp. 67–110). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Olsson, D., & Gericke, N. (2016). The adolescent dip in students' sustainability consciousness: Implications for education for sustainable development. *The Journal of Environmental Education*, *47*(1), 35–51.

Olsson, D., & Gericke, N. (2017). The effect of gender on students' sustainability consciousness: A nationwide Swedish study. *The Journal of Environmental Education*, *48*(5), 357–370.

Pettersson, L. (2014). *Att mötas i tid, rum och tanke: Om ämnesintegration och undervisning för hållbar utveckling, Kulturgeografiska institutionen, Forskarskolan i geografi*. Uppsala: Uppsala universitet.

Rittel, H. W., & Webber, M. (1973). Dilemmas in general theory of planning. *Policy Sciences*, *4*(2), 155–169.

Rodriguez, M. C. (2016). Selected-response item development. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 259–274). New York, NY: Routledge.

Sandell, K., Öhman, J., & Östman, L. (2003). *Miljödidaktik. Naturen, skolan och demokratin*. Lund: Studentlitteratur.

Schulz-Heidorf, K., & Støle, H. (2018). Gender differences in Norwegian PIRLS 2016 and ePIRLS 2016 results at test mode, text and item format level. *Nordic Journal of Literacy Research*, *4*(1), 167–183. .

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation*, *22*(1), 4.

Skolverket [Swedish National Agency for Education]. (2011). *Curriculum for the compulsory school grades*, 7–9. Stockholm, Sweden.

Skolverket [Swedish National Agency for Education]. (2018). PISA 2018: 15-åringars kunskaper i läsförståelse, matematik och naturvetenskap, Internationella studier 487, Elanders, Stockholm. https://www.skolverket.se/getFile?file=5347.

Statistics Sweden. (2017). Statistisk uppföljning av Agenda 2030. ISSN 1654-0743 (Online) URN:NBN:SE:SCB-2017-X41BR1701_pdf, Stockholm, Sweden.

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., … Upmeier zu Belzen, A. (2016). Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education*, *41*(5), 721–732.

Sund, P. (2008). *Att urskilja selektiva traditioner i miljöundervisningens socialisationsinnehåll – implikationer för undervisning för hållbar utveckling* Mälardalen University Press Dissertations, Arkitektkopia, Västerås Nr. 63

Sveriges Regering [Government of Sweden]. (2017). Nationella prov – rättvisa, likvärdiga, digitala. Regeringens proposition Prop. 2017/18:14. Utbildningsdepartementet, Stockholm.

Tennant, A., & Pallant, J. F. (2007). DIF matters: A practical approach to test if differential item functioning makes a difference. *Rasch Measurement Transactions*, *20*(4), 1082–1084.

Thornes, J. E. (2004). The visual turn and geography. (Response to Rose 2003 Intervention). *Antipode*, *36*(5), 787–794.

Tomaszewski, B., Vodacek, A., Parody, R., & Holt, N. (2015). Spatial thinking ability assessment in Rwandan secondary schools: Baseline results. *Journal of Geography*, *114*(2), 39–48.

Torbjörnsson, T., & Molin, L. (2014). Who is solidary? A study of Swedish students' attitudes towards solidarity as an aspect of sustainable development. *International Research in Geographical and Environmental Education*, *23*(3), 259–277.

United Nations [UN]. (2002). *Report of the world summit on sustainable development*. New York: United Nations.

United Nations Educational, Scientific and Cultural Organisation [UNESCO]. (2014). *UNESCO roadmap for implementation the global action programme on education for sustainable development*. Paris, France: UNESCO.

van Hek, M., Buchmann, C., & Kraaykamp, G. (2019). Educational systems and gender differences in reading: A comparative multilevel analysis. *European Sociological Review*, *35*(2), 169–186.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, *101*(4), 817–835.

Wiek, A., Withycombe, L., & Redman, C. L. (2011). Key competencies in sustainability: A reference framework for academic program development. *Sustainability Science*, *6*(2), 203–218.

Wikström, C. (2013). *Konsten att göra bra prov – vad lärare behöver veta om kunskapsmätning*. Lettland: Natur & Kultur.

Zieky, M. (2003). *A DIF primer*. Princeton, NJ: Educational Testing Service. Retrieved August 18, 2020, from https://www.ets.org/s/praxis/pdf/dif_primer.pdf.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)* (pp. 1–57). Ottawa: National Defense Headquarters.