ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
Araştırma Makalesi (Research Article); 2022; 13(1); 40-53

# Analyzing the Effects of Test, Student, and School Predictors on Science Achievement: An Explanatory IRT Modeling Approach

Serap BÜYÜKKIDIK *  Okan BULUT **

**Abstract**

This study aimed to investigate the impact of item features (i.e., content domain), student characteristics (i.e., gender), and school variables (i.e., school type) on students' responses to a nationwide, large-scale assessment in Turkey. The sample consisted of 7507 students who participated in the 2016 administration of the Transition from Primary to Secondary Education Exam (TPSEE, referred to as "TEOG" in Turkey). Explanatory item response modeling was used for analyzing the effects of content domain, gender, school type, and their interactions on students' responses to the science items on the exam. Five explanatory models were constructed to examine the effects of the item, student, and school variables sequentially. Results indicated that female students were more likely to answer the items correctly than male students. Also, students from private schools performed better than students from public schools. In terms of content, the biology items appeared to be significantly easier than the physics items. All interactions between the predictors were significant except for the Gender x School Type and Content x Gender x School Type interactions. The interactions between the predictors suggested that test developers, teachers, and stakeholders should be aware of potential item-level bias occurring in the science items due to complex interactions among the items, students, and schools characteristics.

*Keywords:* Explanatory item response modeling, science, school type, gender, content

## Introduction

Being placed in a good high school is often considered an educational pathway to entering a good university and finding a good profession (Carnevale et al., 2018). Every student graduating from primary education is also entitled to attend high school education in Turkey. The transition of students from primary education to secondary (i.e., high school) education in Turkey is generally subject to a nationwide, large-scale assessment. Turkey has a long history of standardized, large-scale assessments with its frequently changing examination system over the years. Since 1999, a central examination system has been applied for entrance to high schools and the High School Entrance Examination (HSEE, abbreviated as "LGS" in Turkish) had been used from 2000 to 2004. After HSEE, students were selected with the High School Institutions Selection and Placement Exam (HSISPE, abbreviated as "OKS" in Turkish), the Level Determination Exam (abbreviated as "SBS" in Turkish), and the Transition from Primary to Secondary Education Exam (TPSEE, referred to as "TEOG" in Turkey), respectively. After the 2013-2014 academic year, the Ministry of National Education (MoNE) in Turkey replaced Level Determination Exam (known as "SBS") with the TPSEE. As a standardized test, TPSEE consisted of 120 multiple-choice items focusing on several subject areas such as Turkish, mathematics, science, and social studies. TPSEE scores were used in calculating the scores required for high school placement in the 2016-2017 academic year (MoNE, 2015). The Transition System to High Schools (referred to as "LGS") has been implemented by the MoNE in Turkey, starting from the 2017-2018 academic year.

In Turkey, where the examination systems change so rapidly and so often, the standardized exams must be valid to serve the purpose of the test, and they must be fair for all students, regardless of the

* Assist. Prof., Sinop University, Faculty of Education, Sinop-Turkey, sbuyukkidik@gmail.com, ORCID ID: 0000-0003-4335-2949
** Assoc. Prof., University of Alberta, Centre for Research in Applied Measurement and Evaluation, Edmonton-Canada, bulut@ualberta.ca, ORCID ID: 0000-0001-5853-1267

To cite this article:
Büyükkıdık, S., & Bulut, O. (2022). Analyzing the effects of test, student, and school predictors on science achievement: An explanatory IRT modeling approach. *Journal of Measurement and Evaluation in Education and Psychology, 13*(1), 40-53. https://doi.org/10.21031/epod.1013784

*Received: 23.10.2021*
*Accepted: 5.01.2022*

gender and socio-economic status of the students. It is essential to examine national-level exams such as the TPSEE and international practices such as the Program for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), and the Trends in International Mathematics and Science Study (TIMSS) in terms of gender, school type and content domain for collecting validity argument. Individuals are placed in high school in line with the scores obtained in the TPSEE. Important decisions that will affect the lives of individuals, such as placement and selection, have been taken through central exams for many years in Turkey, as mentioned. It is one of the primary duties of test developers to prepare fair items and response options on the exams. Ensuring fairness and equality of opportunity in education are two major issues in the Turkish National Education System. Berberoğlu and Kalender (2005) revealed the differences in achievement between schools as a result of their analysis using PISA and national exams, and they stated that these findings contradict the principle of equality of opportunities in education and that all individuals do not benefit from these opportunities equally. Organisation for Economic Co-operation and Development (OECD, 2021) concluded similar results, emphasizing that students from socioeconomically disadvantaged families are less likely to perform well in school, and it is also noted that they do not have a study environment at home or cannot receive support from their parents for their studies. This socio-economic gap is the largest in Turkey when comparing other OECD countries (OECD, 2021).

Another important issue of fairness is to prevent gender bias in items. The items should not contain bias in favor of a group to ensure fairness, especially in large-scale assessments. On the other hand, the first step in bias studies is differential item function (DIF) studies. Previous research on exams mostly involves DIF studies that vary according to school type and/or gender (Ayva Yörü & Atar, 2019; Kalaycioğlu & Berberoğlu, 2010). DIF determination is one of the validity arguments that reveal the group differences in items for individuals with similar ability levels that support the inferences obtained from test scores in test development (American Educational Research Association [AERA] et al., 2014). Explanatory item response modeling (EIRM) can also be used in research to reveal differences in students' performance in terms of gender, school type, and content domain and to reveal the psychometric properties of national and international exams.

The purpose of this study was to examine the impact of three types of covariates on students' performance in the TPSEE: content domain as an item-level predictor, gender as a student-level predictor, and school type (i.e., public vs. private) as a school-level predictor. Explanatory item response modeling was used for evaluating the main effects of these covariates, as well as their interactions, in the science test of the TPSEE. Results of this study are expected to shed light on the complex relationship between item-level, student-level, and school-level predictors.

## Literature Review

Individual characteristics and item properties can affect exam performance (Kan et al., 2018; Liou & Bulut, 2020; Liu & Wilson, 2009). For many years, it has been demonstrated that individual characteristics such as gender (Hyde, 2005; Hyde & Linn, 2006; Legewie & DiPrete, 2014; Quinn & Cooc, 2015; Reilly et al., 2015; Sinnes & Løken, 2014), the type of school (e.g., Berberoğlu & Kalender, 2005; Quinn & Cooc, 2015; Zhang & Campbell, 2015), and item properties such as content domain in the test (see Mullis et al., 2020) and their interaction (e.g., Bell, 2001; Burkam et al., 1997; Kalaycioğlu & Berberoğlu, 2010; Lee & Burkam, 1996; Young & Fraser, 1994) can be influential on students' performance in science assessments.

## Student Characteristics

Gender has an undeniably important role in individuals' lives (Legewie & DiPrete, 2014). Due to the nature of human beings, it is usual to have psychological, anatomical, and behavioral differences between males and females (Ober et al., 2008). There is a continuing interest by researchers in investigating the role of gender in determining student performance (Burkam et al., 1997; Meinck & Brese, 2019). Since the First International Science Study (FISS) in the 1970s, international practices

have been trying to reveal differences in science achievement by gender. In addition to uncovering gender differences in achievement, it is important to discover and understand the reasons for these differences. There are three important conceptual perspectives in gender-based research in mathematics and science, according to Lee and Burkam (1996). The first factor is the individual perspective like self-perception, learning ability, values, attitudes, interests, etc. The second is the environmental perspective, which includes internship or mentoring opportunities and social support, classroom dynamics, and similar factors. The third is an interactionist perspective that typically explores multivariate causal models that combine intrinsic and extrinsic forces (Lee & Burkam, 1996).

The reasons for gender differences in achievement can be examined in two ways as those examining the reason from the student dimension, those examining the reasons from the assessment dimension. In the first category, psychological, cultural, social, psycho-bio-social causes are presented with relational and causal models. In the second category, item characteristics that differentiate student performance according to gender are discussed (Liu & Wilson, 2009). Eriksson et al. (2020) examined the reasons in the first category. They stated that gender differences in achievement might differ according to courses and countries. Researchers have revealed the relationship between gender egalitarian values and gender differences in academic achievement. Their research focused on the role of gender-egalitarian values rather than gender differences in opportunities as in other studies. In addition, researchers have revealed that cultural values such as gender egalitarian values play an essential role in reducing gender gaps in academic achievement.

Hyde (2005, p. 581) proposed the "gender similarities hypothesis," which suggests that males and females are generally the same but different in psychological variables, supported by 46 meta-analyses. It is necessary to examine the gender gap issue in mathematics and science literacy and to take important steps to address the underrepresentation of females in science, technology, engineering, and math (STEM) (Reilly et al., 2015). Reilly et al. (2015) have a small but stable average difference in mathematics and science achievement of 12th-grade male students compared to female students in the study investigating gender differences in science and mathematics in the National Assessment of Educational Progress data between 1990 and 2011 in two decades, with an effect size of $d$ = .10 and .13, respectively. Quinn and Cooc (2015) found a significant gap between genders in favor of boys ($d$ = .23) in science at 3rd grade, which decreases slightly by 8th grade using data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999 (ECLS-K).

However, various studies demonstrated that the views of the peers, the roles of teachers, stereotypes, hegemonic cultural beliefs about gender and local interactions, gender segregation of extracurricular activities, and the sociocultural environment affect students' orientation towards STEM fields and science achievement in the graduated school (Legewie & DiPrete, 2014). Several meta-analyses and studies revealed that gender differences have declined recently and are negligible in science achievement, but psychological and sociological perception can be effective in success (Hyde 2005; Hyde & Linn, 2006; Legewie & DiPrete, 2014; Sinnes & Løken, 2014). These findings supported that gender and perception of gender roles in graduated school is a remarkable factor for science achievement and orientation. In addition to studies that reveal gender differences, there are also studies of differential item function (DIF) that have contrasting results related to gender at the item level (e.g., Ayva Yörü & Atar, 2019; Gierl et al., 1999; Kalaycioğlu & Berberoğlu, 2010). It is expected for test fairness that individuals with similar ability levels in terms of the measured feature regardless of content domain show similar performance. Individuals at the same ability level in two different groups, such as gender or school type, will differ in the correct response probability for items showing DIF (AERA et al., 2014; Hambleton et al., 1991).

Another predictor used in this research is school characteristics. Understanding the impact of school characteristics on learning and achievement is significant because public policies affect access to public schools and the quality of schools as well as private school fees and scholarships (Newhouse & Beegle, 2006). Underlying the "school-choice movement" is the belief that while private schools respond to rivalry and excel in providing educational services, this is not the case in public schools (Figlio & Stone, 1997, p. 3). Therefore, private school students routinely are more likely to perform higher on standardized tests than their peers who attend public schools. Going to a private school can

be an acceptable indicator of the good socio-economic status in Turkey. Turkey is in the first place among OECD countries in affecting the socio-economic status of families in the success of students in education life (OECD, 2021). According to Quinn and Cooc (2015), students with a high socio-economic background tend to perform better than their peers with a low socio-economic status for various reasons. The OECD attributed these reasons to the fact that students from families with low socio-economic status are less likely to have digital learning tools, do not have a study environment in their home, or cannot receive support from their parents for their lessons (OECD, 2021).

Researchers investigated the impact of the socio-economic gap on students' science achievement since elementary school (e.g., Quinn & Cooc, 2015; Zhang & Campbell, 2015). Zhang and Campbell (2015) employed a hierarchical linear model analysis with 9943 8th grade students and 343 middle schools in 6 provinces and 2084 teachers and revealed that the large socio-economic status (SES) gaps in science achievement emerged at the socio-economic level measured at the school level, while the moderate SES gap emerged when SES was measured at the student level. They also mentioned that schools with relatively high socio-economic levels are more likely to have high-quality teachers than schools with lower socio-economic levels. Newhouse and Beegle (2006) concluded that public junior secondary schools were more effective in terms of cognitive abilities and achievement in the national exam than their private equivalents in their research on how school type affects the success of secondary school students. Young and Fraser (1994) attempted to reveal the school x gender interaction with a Hierarchical Linear Model. They found that the influence of the school was greater than the gender effect on physics achievement. In the literature, no completely similar research has been found that demonstrates the interaction of gender and school type with EIRM.

## Item Properties

Liou and Bulut (2020) suggested the use of EIRM with important item properties for testing such as content domains such as biology and physics. Considering Turkey's TIMSS 2019 average science achievement in each of the three content domains, the physical sciences had the highest average while life sciences had the lowest average (Mullis et al., 2020). Gender differences and inequalities in science achievement by content domain from past to present have been one of the research topics for researchers and associations (e.g., from International Association for the Evaluation of Educational Achievement [IEA]'s FISS to TIMSS 2019). Several studies examined gender differences and the differential item function according to gender in the science achievement by the content domain (e.g., Bell, 2001; Burkam et al., 1997; Kalaycioğlu & Berberoğlu, 2010; Lee & Burkam, 1996). When all these studies were examined, it was concluded that female students performed higher than males in life sciences such as biology, while it came out that the situation was in favor of males in physics (Bell, 2001; Kalaycioğlu & Berberoğlu, 2010; Lee & Burkam, 1996; Mullis et al., 2020). Burkam et al. (1997) found out that there was an above-moderate advantage in favor of boys in physical sciences, while the life science test performance of girls was relatively higher among the less able students.

In the cross-cultural research conducted on 17 countries by the IEA, a performance difference was found in favor of males in science achievement in all age groups, in the FISS, the biggest gender gap in science is in favor of male in physics, the least, it has been revealed to be in biology (IEA, 1988). The results of the TIMSS 2019, the most recent study conducted by the IEA, showed that almost half of the participating countries have gender equality in mathematics and science achievement. Given the TIMSS 2019 average science achievement, there was gender equity in average science achievement in 33 countries, whereas fourth-grade girls had higher average achievement than boys in 18 countries, and boys had the edge in science achievement over girls in seven countries. Eight-grade girls had a considerable advantage in biology and chemistry, whereas boys had superiority in physics and Earth science. When Turkey TIMSS 2019 data were analyzed, average science achievement had a 10-point difference among males and females in the 8th grade, but this difference was not significant. When it was examined the difference between the genders in the biology and chemistry content domain in the 8th-grade students in Turkey, girls showed higher performance than boys, and this difference was significant. In the field of physics, there was a 2-points difference in favor of girls, which was not

significant (Mullis et al., 2020). Young and Fraser (1992) examined gender differences in physics achievement with multilevel analysis in terms of socio-educational level, school type (government, Catholic and independent), and sex composition of the school (single-sex and coeducational). They found out that there was a difference in favor of males in all school types two decades ago.

Traditional item response theory (IRT) models reveal information about items for the selected IRT model by considering the levels of respondents' characteristics such as achievement, cognitive ability, etc., with item difficulty, discrimination, and pseudo guessing parameters. Traditional IRT models do not allow analysis by including items and respondent attributes based on the design or theory behind the measuring tool. This stage is a step that should not be neglected for test developers as it provides important information about the measured structure (AERA et al., 2014). Explanatory item response theory models eliminate these limitations of traditional IRT.

### Purpose of the Study

This study aims to contribute to the literature by examining the effects of student characteristics, item properties, and their interactions together on science achievement scores in the TPSEE using the EIRM framework. The data included students' responses in the 2016 administration of the TPSEE to the science subtest covering two content domains (physics and biology), gender (female or male), and school type (public or private). Using the TPSEE data, the following research questions were addressed:

1. To what extent did the students' performance vary by gender and school type (public, private) in the TPSEE science assessment?
2. To what extent did the students' performance vary by the content domain (biology, physics), gender, and school type in the TPSEE science assessment?
3. Which content domains were easier for female and male students to get higher scores in the TPSEE science assessment?
4. To what extent did the students' performance vary by content, school type interaction, and gender effect in the TPSEE science assessment?
5. To what extent did the student performance vary based on the interactions of gender, content domain, and school type in the TPSEE science assessment?

## Method

As a cross-sectional, explanatory research study, this study aimed to examine the effects of student characteristics, school variables, and item features on student performance in the science subtest of the TPSEE.

### Participants

The sample of this study consists of 7507 students who were randomly selected from the 8th-grade students who participated in the TPSEE in November 2016. The answers of the students who took the "booklet A" were used in the research. Table 1 presents the descriptive statistics.

**Table 1**
*Descriptive Statistics for Sample*

| Gender | Frequency | Percent | School Type | Frequency | Percent |
|---|---|---|---|---|---|
| Female | 3879 | 51.7 | Public School | 5481 | 73 |
| Male | 3628 | 48.3 | Private School | 2026 | 27 |
| Total | 7507 | 100.0 | Total | 7507 | 100 |

The number of female students ($n$ = 3879) was slightly higher than that of male students ($n$ = 3628). 73% of the participants ($n$ = 5481) were a public school students, while the remaining 27% ($n$ = 2026) were private school students.

### Data Collection Instruments

Response data from the 2016 administration of the TPSEE were used in the current study. The data were obtained from the MoNE in Turkey. TPSEE consisted of several subtests focusing on Turkish, mathematics, science, social studies, religious culture and moral knowledge, and foreign language. The student achievement in the science subtest of TPSEE was examined in terms of content domain, gender, and school type to address the research questions of this study. The science subtest measured the construct of competency in science based on 20 items. Each of the biology and physics sections in the science subtest included ten multiple-choice questions. The first ten items were items related to biology, while the last ten items were related to physics. The response data from the science subtest of TPSEE was analyzed using the Rasch model. The item characteristic curves and item information functions based on the Rasch model are shown in Figures 1 and 2, respectively.

**Figure 1**

*Item Characteristic Curves (ICCs)*

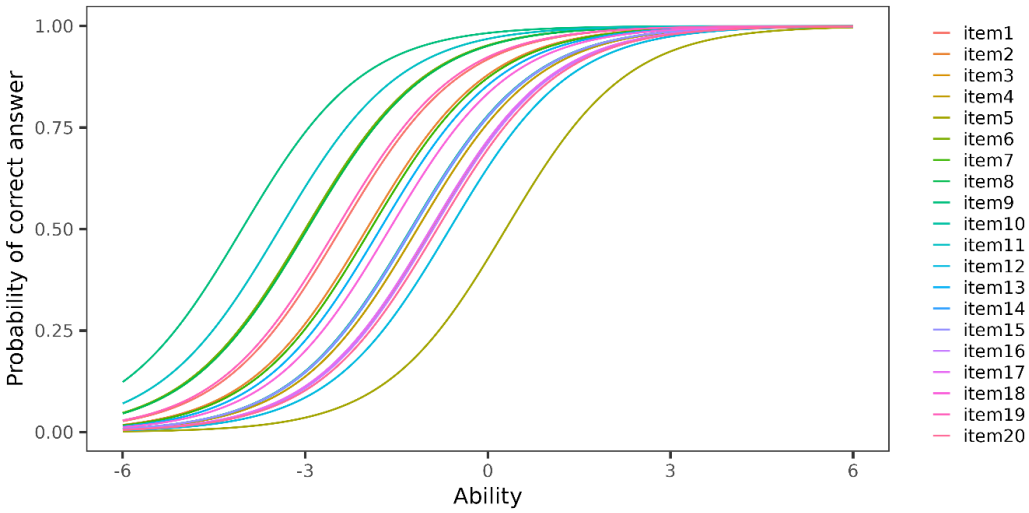



**Figure 2**

*Item Information Curves (IIFs)*

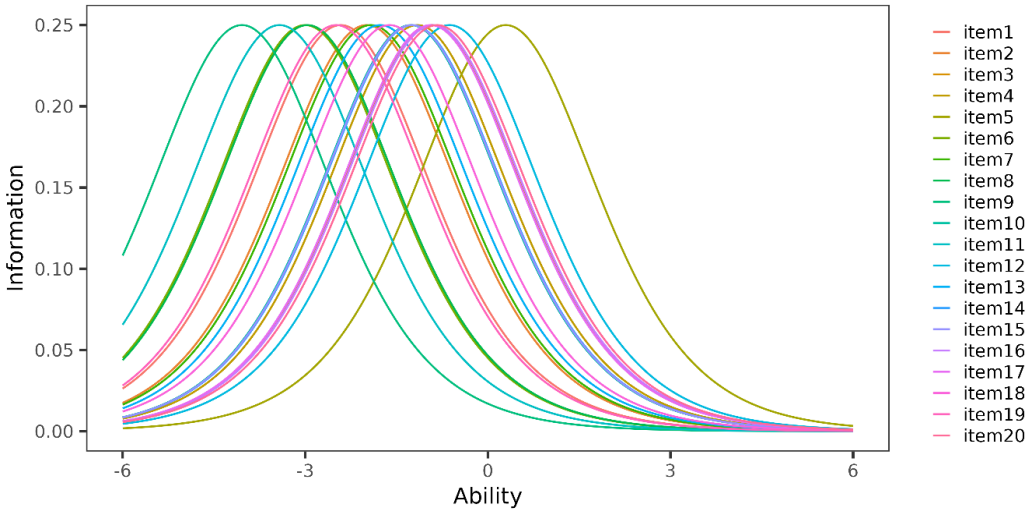


Figure 3 shows the test information function for the science subtest (i.e., the sum of the individual item information functions across the 20 items). As the amount of information obtained from the test information function increases, the standard error of the ability estimation decreases. The science subtest was particularly informative for the range of θ = -3 and θ = +1. The amount of test information was considerably low, especially for the student at a higher ability level.

**Figure 3**

*Test Information Curve and Standard Error (SE)*

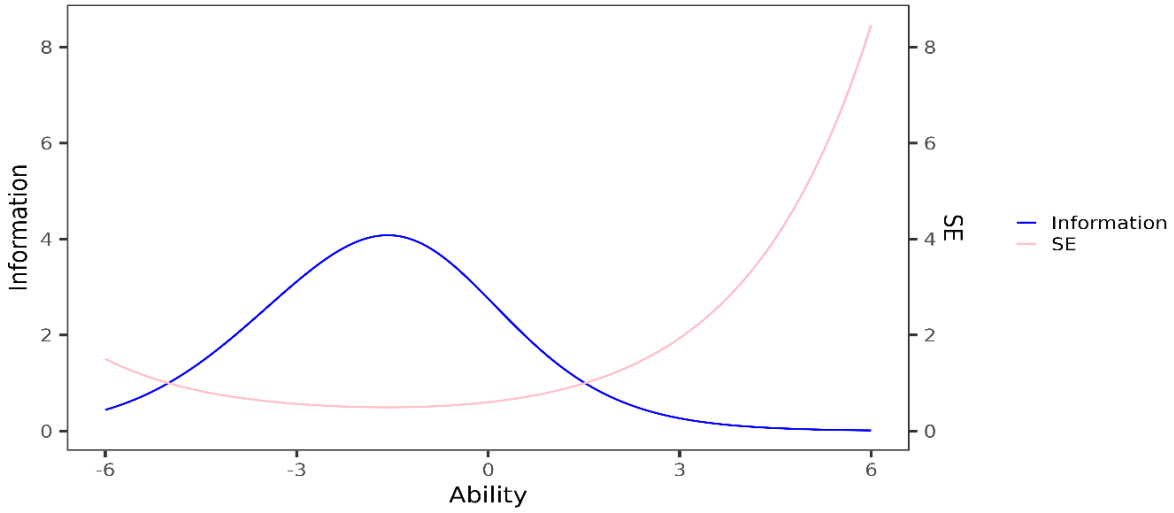
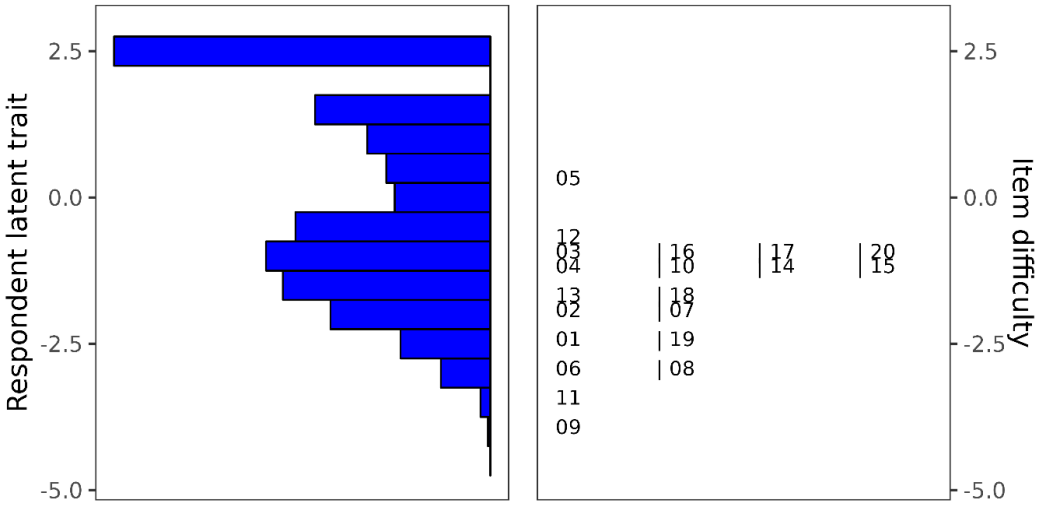



Figure 4 shows the Wright map for the science subtest. The left-hand side of the Wright map shows the histogram of the latent trait distribution, while the right-hand side of the Wright map shows the estimates of the difficulty of each item. The fifth item (at the top) was the most difficult, while the ninth item (at the bottom) was the easiest.

**Figure 4**

*The Wright Map*



## Data Analysis

The first part of the data analysis focused on checking the assumptions of IRT for the science subtest of TPSEE. Parallel analysis was performed for testing unidimensionality, and Yen's Q3 test (Yen, 1984) was used for checking local item independence. The subtest indicated a unidimensional (i.e., one-factor) structure. The lowest factor loading in the one-dimensional structure was .414. The residual correlations for each item pair were below .20. Furthermore, the reliability coefficient of the subtest was KR-20 = .879. After collecting validity and reliability evidence, item characteristic curves and item information functions were examined based on the Rasch model. The "ShinyItemAnalysis" package (Martinková & Drabinová, 2018) was used to obtain the estimates of item parameters and students' ability. In the second part, EIRM was used to examine the effects of predictors at the item, student, and school levels. EIRM is a framework that allows measuring covariates in item sets, student groups, or interactions between item sets and student groups (De Boeck & Wilson, 2004). The main function of EIRM is (1) the odds of accuracy at the item level, (2) individual differences in item-level accuracy, (3) how much of the variance of item level accuracy is due to differences between items rather than interpersonal differences, (4) the chosen person, item predictors and interactions to reveal the explanation of variances (Petscher et al., 2020). Rather than calculating the descriptive effects on the student's feature level or item difficulty, EIRM allows obtaining information by taking into account the explanatory variables with the responses. Generalized linear mixed models (GLMMs) of larger class' models of traditional IRT models can be formulated in EIRM. GLMMs is a function of EIRM when the model includes an item covariate, a person covariate, or a person-by-item covariate (De Boeck & Wilson, 2004; Rijmen et al., 2003; Wilson et al., 2008).

Four explanatory IRT models were estimated using the "lme4" (Bates et al., 2015) package in R (R Core Team, 2021). Table 2 presents the formulation of the four models used in the current study. As the baseline model, Model 0 (Rasch model) did not consist of any predictors. Only gender and school type variables were taken into account in Model 1, while variables of content domain, gender, and school type were covered in Model 2. Content and school type interaction and gender variable were considered in Model 3. In Model 4, the interactions of all variables (content, school type, gender) were examined together.

**Table 2**
*Formulas of the Rasch and Explanatory IRT Models*

| Model | The formula used in lme4 |
|---|---|
| Model 0 (Rasch) | response ~ -1 + item + (1 \| id) |
| Model 1 | response ~ -1 + item + gender + school + (1 \| id) |
| Model 2 | response ~ -1 + content + gender + school + (1 \| id) |
| Model 3 | response ~ -1 + content*school + gender + (1 \| id) |
| Model 4 | response ~ -1 + content*gender*school + (1 \| id) |

## Results

The results for fit indices were shown in Table 4, and it was observed that Model 1 had the best model fit indices based on the AIC, BIC, and log-likelihood, deviance values, suggesting that using gender and school explained the response data better than the other models. Table 3 shows the item easiness parameter and its standard error for Rasch Model without covariates in EIRM and Model 1.

When we looked at Table 3, the fixed effects for estimated values refer to the estimated item (easiness) parameters for the items. Based on the findings, we found that item9 is the easiest item and item5 is the most difficult item based on Rasch Model without covariates and Model 1. Table 4 presents the estimate values for the five EIRM used in the research, their standard error and significance were given.

**Table 3**

*Item Parameters in Respect to Rasch Model Without Covariates in EIRM and Model 1*

| | Model 0 | | | Model 1 | | |
|---|---|---|---|---|---|---|
| Item | Estimate | SE | Z value | Estimate | SE | Z value |
| Item 1 | 2.412 | 0.041 | 58.584* | 1.720 | 0.046 | 37.758* |
| Item 2 | 1.986 | 0.039 | 51.167* | 1.294 | 0.043 | 29.758* |
| Item 3 | 0.917 | 0.036 | 25.539* | 0.223 | 0.041 | 5.428* |
| Item 4 | 1.156 | 0.036 | 31.871* | 0.463 | 0.041 | 11.182* |
| Item 5 | -0.294 | 0.036 | -8.227* | -0.996 | 0.042 | -23.923* |
| Item 6 | 2.994 | 0.046 | 65.068* | 2.301 | 0.050 | 46.106* |
| Item 7 | 1.920 | 0.039 | 49.839* | 1.228 | 0.043 | 28.401* |
| Item 8 | 2.959 | 0.046 | 64.805* | 2.265 | 0.050 | 45.699* |
| Item 9 | 4.030 | 0.062 | 65.501* | 3.333 | 0.064 | 51.776* |
| Item 10 | 1.264 | 0.036 | 34.651* | 0.571 | 0.042 | 13.747* |
| Item 11 | 3.415 | 0.051 | 66.896* | 2.720 | 0.055 | 49.887* |
| Item 12 | 0.626 | 0.036 | 17.557* | -0.069 | 0.041 | -1.685 |
| Item 13 | 1.764 | 0.038 | 46.549* | 1.072 | 0.043 | 25.106* |
| Item 14 | 1.237 | 0.036 | 33.963* | 0.544 | 0.042 | 13.110* |
| Item 15 | 1.250 | 0.036 | 34.307* | 0.558 | 0.042 | 13.429* |
| Item 16 | 0.896 | 0.036 | 24.978* | 0.203 | 0.041 | 4.924* |
| Item 17 | 0.938 | 0.036 | 26.098* | 0.244 | 0.041 | 5.932* |
| Item 18 | 1.611 | 0.037 | 43.095* | 0.919 | 0.042 | 21.731* |
| Item 19 | 2.486 | 0.042 | 59.652* | 1.793 | 0.046 | 38.998* |
| Item 20 | 0.831 | 0.036 | 23.201* | 0.137 | 0.041 | 3.330* |

* $p < .001$

**Table 4**

Results from Five EIRMs

| Predictor | Model 0 | Model 1 β (SE) | Model 2 β (SE) | Model 3 β (SE) | Model 4 β (SE) |
|---|---|---|---|---|---|
| Gender Female (GenderF) | | 0.235 (0.040)* | 0.202 (0.035)* | 0.202 (0.035)* | 0.465 (0.042)* |
| School Private (SchoolP) | | 2.238 (0.049)* | 1.955 (0.042)* | 2.085 (0.048)* | 2.072 (0.064)* |
| Content (Biology) | | | 1.014 (0.027)* | 0.997 (0.027)* | 0.886 (0.029)* |
| Content (Physics) (ContentP) | | | 0.670 (0.027)* | 0.715 (0.027)* | 0.811 (0.029)* |
| ContentP x SchoolP | | | | -0.230 (0.039)* | -0.212 (0.052)* |
| Content P x GenderF | | | | | -0.442 (0.028)* |
| GenderF x SchoolP | | | | | -0.080 (0.097) |
| ContentP x GenderF x SchoolP | | | | | 0.009 (0.079) |
| Model Fit | | | | | |
| AIC | 133180.1 | 131169.6 | 148069.6 | 148036.9 | 147780.7 |
| BIC | 133388.4 | 131397.8 | 148119.2 | 148096.4 | 147869.9 |
| Log Likelihood | -66569.1 | -65561.8 | -74029.8 | -74012.5 | -73881.3 |
| Deviance | 133138.1 | 131123.6 | 148059.6 | 148024.9 | 147762.7 |

*Note*. Male was the reference category for Gender. Public school was the reference category for School. Biology was the reference category for Content in the interactions.

* $p < .001$.

The fixed effects of the gender and school type were expressed as two estimates in Model 1: 0.23 for females and 2.24 for private schools. The effects for gender and school type were statistically significant ($\beta_{Female} = 0.23$, $\beta_{Private} = 2.24$, $p < .001$). The magnitude of the effect was 0.23logits = 1.26 odds ratios for females. Odds ratios larger than 1 indicate higher likelihood, whereas odds ratios smaller than 1 indicate lower likelihood (Liou & Bulut, 2020). This result indicated that 1.26 times more likely to get a higher score if the gender is female. The magnitude of the effect was 2.24logits =

9.39 odds ratios for private schools. This result indicated that 9.39 times more likely to get a higher score if the school type is private.

Table 4 shows the results for Model 2 where content, gender, and school type were used together. The effects for gender and school type were statistically significant ($\beta_{Female} = 0.20$, $\beta_{Private} = 1.95$, $p < .001$). The magnitude of the effect was 0.20logits = 1.22 odds ratios for females. This result indicated that 1.22 times more likely to get a higher score if the gender is female. The magnitude of the effect was 1.96logits = 7.10 odds ratios for private schools. This result indicated that 7.10 times more likely to get a higher score if the school type is private similar to Model 1. The regression coefficients for biology and physics were all statistically significant in Model 2 ($\beta_{Biology} = 1.01$, $\beta_{Physics} = 0.70$, $p < .001$). The items associated with the biology content domain ($\beta_{Biology} = 1.01$) were easier than the physics content domains.

Model 3 in Table 4 shows the estimated item easiness for the gender effect and content x school type interactions. All interaction effects were significant ($p < .001$). Female students get more likely higher scores than male students, which is similar to Model 1 and Model 2. If the content domain is physics, it is less likely to get a higher score than biology similar to Model 2. The physics content domain was more difficult for the students from private schools to get higher scores ($e^{-0.23} = 0.80$).

Table 4 shows the estimated item easiness for the content x gender x school interactions in Model 4. All interaction effects were significant ($p < .001$) except for the Gender x School and Content x Gender x School interactions. The magnitude of the effects was 1.60 and 7.93 for female and private schools, respectively. The magnitude of the effects was 2.44 and 2.25 for biology and physics, respectively. If the content domain is physics, it is less likely to get a higher score than biology, which is similar to Model 2 and Model 3. The physics content domain was more difficult for the female students to get higher scores ($e^{-0.44} = 0.64$). Lastly, when compared to public school students, private school students were less successful in the physics content domain ($e^{-0.21} = 0.81$).

## Discussion and Conclusion

This study aimed to explain the differences in item difficulty and student performance in the science subtest of TPSEE using explanatory IRT models with an item feature (content domain), a student characteristic (gender), a school-related predictor (private or public schools), and their interactions. Furthermore, the difficulty of the items in the science subtest was examined using the Rasch model without any covariates (Null model). Five Explanatory IRT models were used in this study. The results for model 1 indicated that the probability of obtaining a high score was higher for private schools compared to state schools. It was concluded that the performance of females in the science test was higher than males' performance. In Model 2, it was found that the items in the biology content domain were easier than the items in the physics content domain. The interaction effect was found to be significant in Model 3, and parallel results were obtained with Model 1 and Model 2. In Model 4, half of the interactions between the predictors were significant.

The result showed that female students outperformed male students in the science test. Previous studies on gender-based achievement differences in science have been inconclusive (Hyde, 2005; Hyde & Linn, 2006; Wang et al., 2013). In an early study, Young and Fraser (1994) argued that the underrepresentation of females in science was typically attributed to their poor performance in science. A similar view has been discussed by other researchers (e.g., Lee & Burkam, 1996; Legewie & DiPrete, 2014; Sinnes & Løken, 2014). Researchers argued that while there is no differentiation in the abilities of individuals of both genders to be successful in science under equal conditions, discriminatory attitudes between both genders lead to differentiation. Other researchers assumed that the differentiation between females and males is within the framework of interests, values, and abilities and that these differences must be addressed and met to reduce gender differences in science education (Sinnes & Løken, 2014). However, recent studies using large-scale assessments indicate an achievement gap in science in favor of female students. For example, in Finland, as one of the most successful countries in international large-scale assessments, 15-year-old female students performed

higher than male students in PISA 2015 (OECD, 2016). Stoet and Geary (2018), in their study with 472242 students, found that female students' science achievement was similar to or better than males in two-third of the countries participating in PISA 2015.

In addition to examining the science subtest by gender, the gender x content domain interaction was also utilized in this current study. In previous studies, gender differences were examined, taking into account the content domain in science (e.g., Bell, 2001; Burkam et al., 1997; Kalaycioğlu & Berberoğlu, 2010; Lee & Burkam, 1996; Mullis et al., 2020). From the FISS to TIMSS 2019, males were more successful in physics, while differences were more in favor of females in biology (see Mullis et al.'s, 2020 report). Bell (2001) found differentiation in physics content in favor of males and biology content in favor of females in the retrieval of declarative knowledge and not the use of procedural knowledge question sections. Bell (2001) related this difference with the differentiation of out-of-school activities according to gender.

The interaction effect of gender x content domain was statistically significant. The physics content domain was more difficult for the female students to get higher scores. This finding shows that while females perform better than males in general science achievement, the physics content field is easier for males, which is compatible with the studies of Bell (2001) and Kalaycioğlu and Berberoğlu (2010) and Lee and Burkam (1996) and TIMSS 2019 results. Kalaycioğlu and Berberoğlu (2010), in their DIF study conducted by University Entrance Examinations in Turkey, revealed that four physics questions showed DIF in favor of males, and a total of six chemistry and biology questions showed DIF in favor of females in the science subtest. When Lee and Burkam (1996) divided the standard tests measuring general science achievement into its life science and physical science domains, they found that males showed an advantage in the physics subtest, while females were relatively advantageous in the field of life science. Burkam et al. (1997) conducted their research with large-scale national longitudinal data, while the moderate advantage at the 8th grade was in favor of males, this advantage expanded in the physical sciences test in the 10th grade. Considering the ability levels of students, average and above-average level males were more successful, while below-average females performed better in life sciences.

The effect for school type was statistically significant. The result indicated that more likely to get a higher score if the school type is private. Young and Fraser (1992) found out that school type was an important indicator of physics achievement in their multilevel study. Young and Fraser (1994) claimed that gender differences in science performance were rarely examined with the interaction of gender and school environment and processes. They stated that when the school effect was ignored, student differences caused biased statistical significance, so gender differences were related to social factors at home and school. Gender differences were higher in some schools, and that the variances between schools were neglected by the researchers. The Hierarchical Linear Model had shown that 10% and 20% of the school effect was effective in physics achievement, and the gender of the students was 3% effective.

In line with the findings of this research, the past studies revealed that the success of private schools was higher than public schools due to socio-economic indicators and other factors (Coleman et al., 1982; Figlio & Stone, 1997). Zhang and Campbell (2015) found that there is a high school-level SES gap in science achievement because schools with high socio-economic levels may have more qualified teachers. Newhouse and Beegle (2006) revealed that public schools had standard deviations difference from 0.17 to 0.3 compared to private schools. This difference in favor of public schools differs from the results of this research. Berberoğlu and Kalender (2005) revealed that there are great differences between school types in terms of learning outcomes. They stated that everyone should benefit from equal learning conditions by reducing the differences in achievement between schools in the social country.

## Limitations and Suggestions

Our research has some limitations. Firstly, causality interpretation cannot be made from the differences regarding item-level and student-level variables. For example, it cannot be inferred that the female gender is a reason for high success in science. Secondly, five models were discussed in the EIRM application; the findings of the research are limited to these models discussed in the study. Thirdly, the research is limited to the findings obtained from the responses of 7507 participants to the 2016 November TPSEE science test. Similar studies can be designed considering all these limitations of the research.

The research results provide information based on empirical and practical evidence for test developers, policymakers, teachers, MoNE in Turkey, organizations such as the student selection and placement center (SSPC, known as ÖSYM in Turkey). Lee and Burkam (1996) and Burkam et al. (1997) stated that increasing experimental, hands-on learning activities and active involvement of students in classrooms encourage gender equality in science achievement, especially in physical sciences. For this purpose, it is recommended to promote such activities for both genders and take steps within the framework of social gender equality for the equal representation of women in STEM careers and gender equity in science. As stated by Hughes (2001), the curriculum should be constructed regardless of gender. While the physical sciences are presented in more masculine content, the biological sciences are more feminine or gender ambiguous. It is natural for males and females to differentiate in their interests (see Ober et al., 2008). Course contents should be prepared in a way that attracts the attention of both genders, and teachers should conduct their lessons by considering the qualifications of each individual.

Instead of revealing the inequalities or differences arising from gender and school type, research can be conducted to contribute to education that eliminates inequalities under unfair conditions. Equality of opportunity should be ensured, and the differences in achievement between private and public schools should be eliminated. Test development and application centers such as the MoNE and SSPC of Turkey should avoid biased questions based on gender and school type in their measurement practices. It is also recommended to base policymakers' activities on scientific research to reduce inequalities and underrepresentation. Educational processes may differ in terms of gender, ethnicity, socio-economic status, etc. It should be rearranged according to inclusive education to include every child fairly regardless of the characteristics.

## Declarations

**Conflict of Interest:** No potential conflict of interest was reported by the authors.

**Ethical Approval:** Secondary data were used in this study. Therefore, ethical approval is not required.

## References

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

Ayva Yörü, F. G., & Atar, H. Y. (2019). Determination of differential item functioning (DIF) according to SIBTEST, Lord's [chi-squared], Raju's area measurement and Breslow-Day methods. *Journal of Pedagogical Research, 3*(3), 139-150. https://doi.org/10.33902/jpr.v3i3.137

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://arxiv.org/pdf/1406.5823.pdf

Bell, J. F. (2001). Investigating gender differences in the science performance of 16-year-old pupils in the UK. *International Journal of Science Education, 23*(5), 469-486. https://doi.org/10.1080/09500690120123

Berberoğlu, G., & Kalender, İ. (2005). Öğrenci başarısının yıllara, okul türlerine, bölgelere göre incelenmesi: ÖSS ve PISA analizi. *Journal of Educational Sciences & Practices, 4*(7), 21-35. http://www.ebuline.com/english/pdfs/7_2.pdf

Burkam, D. T., Lee, V. E., & Smerdon, B. A. (1997). Gender and science learning early in high school: Subject matter and laboratory experiences. *American Educational Research Journal, 34*(2), 297-331. https://doi.org/10.3102/00028312034002297

Carnevale, A. P., Strohl, J., Ridley, N., & Gulish, A. (2018). *Three educational pathways to good jobs: High school, middle skills, and bachelor's degree*. https://1gyhoq479ufd3yna29x7ubjn-wpengine.netdna-ssl.com/wp-content/uploads/3ways-FR.pdf

Coleman, J. S., Hoffer, T., & Kilgore, S. (1982). Cognitive outcomes in public and private schools. *Sociology of Education, 55*(2-3), 65-76. https://www.jstor.org/stable/pdf/2112288.pdf?casa_token=2urqntYyuZQAAAAA:Rj2XcpFpD4Asklsmj_minXEoi7CsxMD1kg7yrb81rVd2wuN_j7zMZu6feBoRaNnN53xFobtwRojV4LkD8cv8WzWUgMWGGMbIBgaVwN2cMWCRp8G0VoI

De Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory item response models*. Springer.

Eriksson, K., Björnstjerna, M., & Vartanova, I. (2020). The relation between gender egalitarian values and gender differences in academic achievement. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.00236

Figlio, D.N., & Stone, J.A. (1997). *School choice and student performance: Are private schools really better?* (Discussion Paper 1141-97). Institute for Research on Poverty, University of Wisconsin-Madison.

Gierl, M., Khaliq, S. N., & Boughton, K. (1999, June). *Gender differential item functioning in mathematics and science: Prevalence and policy implications*. In Annual Meeting of the Canadian Society for the Study of Education, Canada.

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Sage.

Hughes, G. (2001). Exploring the availability of student scientist identities within curriculum discourse: An anti-essentialist approach to gender-inclusive science. *Gender & Education, 13*(3), 275-290. https://doi.org/10.1080/09540250120063562

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*(6), 581-592. https://doi.org/10.1037/0003-066X.60.6.581

Hyde, J. S., & Linn, M. C. (2006). Gender similarities in mathematics and science. *Science, 314*(5799), 599-600. https://doi.org/10.1126/science.1132154

International Association for the Evaluation of Educational Achievement. (1988). *Science achievement in seventeen countries* (A Preliminary Report). Pergamon Press.

Kalaycioğlu, D. B., & Berberoğlu, G. (2010). Differential item functioning analysis of the science and mathematics items in the university entrance examinations in Turkey. *Journal of Psychoeducational Assessment, 29*(5), 467-478. https://doi.org/10.1177/0734282910391623

Kan, A., Bulut, O., & Cormier, D. C. (2018). The impact of item stem format on the dimensional structure of mathematics assessments. *Educational Assessment, 24*(1), 13-32. https://doi.org/10.1080/10627197.2018.1545569

Lee, V. E., & Burkam, D. T. (1996). Gender differences in middle grade science achievement: Subject domain, ability level, and course emphasis. *Science Education, 80*(6), 613-650. https://doi.org/10.1002/(SICI)1098-237X(199611)80:6<613::AID-SCE1>3.0.CO;2-M

Legewie, J., & DiPrete, T. A. (2014). The high school environment and the gender gap in science and engineering. *Sociology of Education, 87*(4), 259-280. https://doi.org/10.1177/0038040714547770

Liou, P. Y., & Bulut, O. (2020). The effects of item format and cognitive domain on students' science performance in TIMSS 2011. *Research in Science Education, 50*(1), 99-121. https://doi.org/10.1007/s11165-017-9682-7

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. https://doi.org/10.1080/08957340902754635

Martinková P., & Drabinová A. (2018) ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal, 10*(2), 503-515. https://doi.org/10.32614/RJ-2018-074

Meinck, S., & Brese, F. (2019). Trends in gender gaps: Using 20 years of evidence from TIMSS. *Large-Scale Assessments in Education, 7*(1), 1-23. https://doi.org/10.1186/s40536-019-0076-3

Ministry of National Education. (2015). *Ortaöğretim kurumlarına geçiş uygulaması tercih ve yerleştirme e-kılavuzu 2015*. http://odsgm.meb.gov.tr/meb_iys_dosyalar/2015_05/28024630_ekilavuz28.05.2015.pdf

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. https://timssandpirls.bc.edu/timss2019/international-results/

Newhouse, D., & Beegle, K. (2006). The effect of school type on academic achievement evidence from indonesia. *Journal of Human Resources, 41*(3), 529-557. https://doi.org/10.3368/jhr.XLI.3.529

Ober, C., Loisel, D. A., & Gilad, Y. (2008). Sex-specific genetic architecture of human disease. *Nature Reviews Genetics, 9*(12), 911-922. https://doi.org/10.1038/nrg2415

Organisation for Economic Co-operation and Development. (2016). *PISA 2015 results: Excellence and equity in education* (Vol. 1). OECD Publishing. https://doi.org/10.1787/9789264266490-en

Organisation for Economic Co-operation and Development. (2021). *Education at a Glance 2021: OECD indicators*. OECD Publishing. https://doi.org/10.1787/b35a14e5-en

Petscher, Y., Compton, D. L., Steacy, L., & Kinnon, H. (2020). Past perspectives and new opportunities for the explanatory item response model. *Annals of Dyslexia, 70*(2), 160-179. https://doi.org/10.1007/s11881-020-00204-y

Quinn, D. M., & Cooc, N. (2015). Science achievement gaps by gender and race/ethnicity in elementary and middle school. *Educational Researcher, 44*(6), 336-346. https://doi.org/10.3102/0013189X15598539

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress Assessments. *Journal of Educational Psychology, 107*(3), 645-662. https://doi.org/10.1037/edu0000012

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*(2), 185-205. https://doi.org/10.1037/1082-989X.8.2.185

Sinnes, A. T., & Løken, M. (2014). Gendered education in a gendered world: looking beyond cosmetic solutions to the gender gap in science. *Cultural Studies of Science Education, 9*(2), 343-364. https://doi.org/10.1007/s11422-012-9433-z

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science, 29*(4), 581-593. https://doi.org/10.1177/0956797617741719

Wang, M.-T., Eccles, J. S., & Kenny, S. (2013). Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics. *Psychological Science, 24*(5), 770-775. https://doi.org/10.1177/0956797612458937

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 91-120). Hogrefe.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125-145. https://doi.org/10.1177/014662168400800201

Young, D. J., & Fraser, B. J. (1994). Gender differences in science achievement: Do school effects make a difference? *Journal of Research in Science Teaching, 31*(8), 857-871. https://doi.org/10.1002/tea.3660310808

Young, D. J., & Fraser, B. J. (1992, April). *Sex differences in science achievement: A multilevel analysis*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco. https://files.eric.ed.gov/fulltext/ED356947.pdf

Zhang, D., & Campbell, T. (2015). An examination of the impact of teacher quality and "opportunity gap" on student science achievement in China. *International Journal of Science and Mathematics Education, 13*(3), 489-513. https://doi.org/10.1007/s10763-013-9491-z