BOĞAZİÇİ UNIVERSITY JOURNAL OF EDUCATION
2024, Vol. 41-3, No. 1
https://doi.org/10.52597/buje.1411656

Investigating Differential Item Functioning of an International Mathematics Competition Items across Gender Groups

Serkan Arıkan[a]

Abstract

Mathematical problem-solving competitions have existed for over a century. Scholars report the gender gap in these competitions. As a result, it is necessary to determine whether any score difference between gender groups is attributable to a genuine difference or is the result of the exam itself. Thus, the current study specifically examined bias in one of the well-known mathematics competitions: the Kangaroo Mathematics competition. Determining the fairness of Kangaroo mathematics competition items across gender groups is crucial for creating accurate comparisons and avoiding unintended construct irrelevant bias. To examine the bias, Differential Item Functioning (DIF) analyses were conducted using Logistic Regression, Mantel-Haenszel, and Item Response Theory Likelihood Ratio Test DIF detection methods. After a series of investigations, out of 336 items, it was concluded that these mathematics items were free of DIF and bias across the gender groups. Further implications were discussed in detail regarding the validity and bias.

Keywords: DIF, bias, mathematical problem-solving competitions, logistic regression, Mantel-Haenszel, Item Response Theory Likelihood Ratio Test

Article info

Received: 29.12.2023
Revised: 04.03.2024
Accepted: 14.03.2024
Published online: 22.04.2024

Introduction

Mathematical problem-solving competitions date back more than a century. Since then, there have been numerous mathematics competitions that have piqued the interest of students. These mathematics competitions are valuable not only for identifying talented students, but also for encouraging and developing mathematical ability, providing both an opportunity to deal with original problem situations, and a new and high-quality material for classroom teaching and math clubs (de Losada & Taylor, 2022). Compared to school-based exams, these competitions have become attractive as they are external and there is no need to worry about the results (de Losada & Taylor, 2022). Similarly, starting in 1991, Kangaroo mathematics competitions have been attracting more than six million students all around the world each year in more than 70 countries. This annual international mathematical game contest, run by a non-profit organization called the Association Kangourou Sans Frontières, aims to increase the popularization of mathematics in schools, support mathematical education, and increase the joy of mathematics by promoting a positive perception towards mathematics in society (Akveld, Caceres-Duque, Geretschläger, 2020). The competition is run on the same day all around the world. There are six categories according to grades (Pre-Ecolier: Grade 1 and 2; Ecolier: Grade 3 and 4, Benjamin: Grade 5 and 6, Cadet: Grade 7 and 8, Junior: Grade 9 and 10 and Student: Grade 11 and 12). Students in each category answer the same questions, but evaluation and comparison are done within each grade level. In order to develop the test, the candidate items are written and submitted in English by member countries and then selected in the Annual Meeting by teachers, mathematicians, and math educators. Then, selected item sets are translated into target languages by country offices (Akveld, Caceres-Duque, & Geretschläger, 2020). Detailed descriptions of the items used in this competition were provided by scholars (Andritsch et al., 2020; Geretschläger & Donner, 2022).

Each country administers their own test, does the scoring of their students, and announces the results. In the Kangaroo Mathematics competition, cross-country score comparisons are not made. Although the Kangaroo Mathematics competition is not a high-stakes exam, students compete fiercely to be successful. As a result, the comparability of test scores across gender groups and the fairness of items for boys and girls are essential issues, as any comparative assessment should be fair to both groups of pupils (International Test Commission, 2001). Scholars are interested in the gender differences in mathematics performance (Hyde, et al. 2008) and some

[a]*Boğaziçi University, Mathematics and Science Education, serkan.arikan1@bogazici.edu.tr, ORCID: 0000-0001-9610-5496*

To cite: Arıkan, S. (2024). Investigating differential item functioning of an international mathematics competition items across gender groups. *Boğaziçi Unıversity Journal of Education*, *41-3*(1), 53-69. https://doi.org/10.52597/buje.1411656

reported males perform far better than females in competitive environments compared to non-competitive environments (Gneezy et al., 2003; Niederle & Vesterlund, 2010). Applebaum and Leikin (2019) reported that the gender gap is task dependent on mathematics competitions. Thus, it is required to identify whether any score difference between gender groups is due to true differences or stems from the test itself. Consequently, assessing the fairness of Kangaroo Mathematics competition items across gender groups is critical for making reliable comparisons and avoiding unintentional construct-irrelevant bias.

When conducting a comparative study or organizing a competition, it is required to distinguish between the impact and the bias. Impact occurs when one group genuinely has more or less ability on the construct of interest. To assess potential bias in scores, Differential Item Functioning (DIF) detection methods are used to evaluate items. DIF occurs and threatens the comparability of test scores of groups when students who have the same ability level on the construct of interest do not have the same probability of answering the item correctly for a single item (Holland & Thayer, 1988; van de Vijver & Leung, 1997; Zumbo, 2007). Examining and presenting evidence that items are free of DIF is necessary for valid score interpretations. Otherwise, if a test has items showing DIF, observed score differences for specific groups could be due to construct-irrelevant variance rather than the true differences in the ability (He & van de Vijver, 2013). DIF is a statistical term and when an item shows DIF, it means that the item might be biased. As a result, to determine whether an item is biased or not, experts examine the item to determine whether the item functions as biased against one group or not (Allalouf et al., 1999; van de Vijver & Leung, 1997). This expert evaluation is done by investigating the possible sources and causes of bias. DIF can stem from a variety of sources, including inadequate translations, unclear original items, low familiarity/appropriateness of the item content in some cultures/groups, the influence of issues specific to a particular culture/groups, such as nuisance factors or connotations associated with the item wording (van de Vijver & Tanzer, 2004), and contextual factors, such as the socioeconomic status and classroom practices (Zumbo & Gelin, 2005).

There has been a lot of research done on identifying biased items (Berrío et al., 2020). DIF studies with real datasets mainly focus on evaluating items of large-scale assessments such as Programme for International Student Assessment **(**PISA), The Trends in International Mathematics and Science Study (TIMSS), and The National Assessment of Educational Progress (NAEP) (Arikan, 2019; Kankaraš & Moors, 2014; Lyons-Thomas et al., 2014; Reynolds et al., 2022; Roberson & Zumbo, 2019; Zwick & Ercikan, 1989), college admission tests (Dorans, 2013; Stark et al., 2004; Wedman, 2018), job application tests (Stark et al., 2004), and licensing examinations (Rubright et al., 2022). The examination of DIF in large-scale and high-stakes tests is required since test scores are utilized to make critical judgments. Today, fairness in classroom assessment is getting attention (Baniasadi et al., 2023).

Present Study

More people are becoming aware of mathematical competitions and Olympiads, such as the TUBITAK Science Olympiads, The International Mathematical Olympiad (IMO), The American Mathematics Competition (AMC) and the Kangaroo Mathematics Competition. Although there are many worldwide mathematics competitions or mathematics Olympiads, there is a lack of research on the psychometric properties of the tests used in these competitions. There is just one research that focus on investigating gender DIF in the American Mathematics Competition contests administered between 2003 and 2007 (Desjarlais, 2009). The results showed that only two of the 125 items were determined to have non-negligible DIF in accordance with the ETS criteria. The Kangaroo competition is one of the most well-known mathematical competitions, with about 6000000 kids competing annually globally. However, to our knowledge, there is no study investigating DIF in Kangaroo Mathematics competition items. Thus, there is a need to evaluate Kangaroo Mathematics items to understand whether any gender differences in test scores may reflect true differences in the mathematics ability or have bias in test items. Therefore, the current study aims to evaluate Kangaroo Mathematics items in terms of gender DIF. To achieve this goal, Kangaroo Mathematics items that are used in the Turkish version of the competition were tested to investigate whether they contain gender DIF or not. The research question of this study is “Do Kangaroo Mathematics items show DIF across gender groups?”

Method

Participants

Kangaroo Mathematics competition is administered to students from grade 1 to grade 12. The current study participants were students who attended Kangaroo Mathematics competition in Turkey in 2022. DIF analyses were conducted for each grade level separately. The number of students in each grade, gender group mean scores, and effect size of the differences are presented in Table 1. Overall, boys performed better than girls; however, the effect size calculations showed that these differences were small according to Cohen's d (Cohen, 1988), except for grade 11. In grade 11, the difference was medium.

Table 1

Descriptive Statistics of Participants

Grade	Number of Students		Mean Scores		Effect Size
	Boys	Girls	Boys	Girls	
Grade 1	3249	2504	7.43	6.94	0.14*
Grade2	5579	4516	10.35	10.02	0.09*
Grade3	7772	6045	8.90	8.33	0.15*
Grade4	5697	4593	11.52	10.73	0.18*
Grade5	6438	5138	11.97	10.62	0.30*
Grade6	4921	3784	15.14	13.66	0.28*
Grade7	3775	3096	10.74	9.74	0.22*
Grade8	2067	1686	14.01	12.54	0.28*
Grade9	1352	1282	17.34	15.62	0.32*
Grade10	1355	1308	17.71	15.91	0.35*
Grade11	761	622	16.96	14.46	0.43*
Grade12	244	175	16.63	15.49	0.18

* $p < 0.05$

Instrument

Kangaroo Mathematics competition have mathematics items that aim to increase student interest in mathematics and promote mathematical thinking. Every year, the competition uses mathematical items that are proposed by each country. Next, representatives from each country gather to choose and determine the final set of items. Following the selection of the items in English, each nation works with its scientific committee to translate and adapt the items. Items that are inappropriate for the national curriculum may be substituted by this committee.

Each grade level consists of 24 or 30 multiple choice questions (from grade 1 to grade 4, 24 items; from grade 5 to grade 12, 30 items) and there are three item categories according to the anticipated difficulty of the items (3-point, 4-point, and 5-point-problems). Grade 1 and 2, grade 3 and 4, grade 5 and 6, grade 7 and 8, grade 9 and 10, grade 11 and grade 12 took the same test but no comparison is made across grade levels. The content dimensions of the test are Numbers, Geometry, Combinatorics, and Algebra. One example item for 7th and 8th graders is provided in Figure 1 (Akveld, Caceres-Duque, Nieto Said & Sánchez Lamoneda, 2020).

Figure 1

Kangaroo Mathematics Example Item

Cadet, 2019-14. Alan, Bella, Claire, Dora, and Erik met at a party and shook hands exactly once with everyone they already knew. Alan shook hands once, Bella shook hands twice, Claire shook hands three times and Dora shook hands four times. How many times did Erik shake hands?

Ⓐ 1; Ⓑ 2; Ⓒ 3; Ⓓ 4; Ⓔ 0.

Data Analysis

There are many methods for DIF detection (Berrío et al., 2020). As with other statistical tests, DIF results can have Type 1 error (false positive). Using multiple DIF detection methods allows researchers a triangulation of the results. Thus, in the current study, three different DIF detection methods were used. To obtain more reliable results, an item flagged as showing DIF in at least two methods was regarded to have DIF across gender groups. The DIF detection methods selected in the current study were logistic regression (LR), Mantel-Haenszel (MH), and the Item Response Theory Likelihood Ratio Test (IRT-LR). LR and MH are based on Classical Test Theory where the ability match is done by total scores and these methods apply parametric and non-parametric models, respectively. IRT-LR is based on Item Response Theory and the ability match is done by latent traits. LR could identify nonuniform DIF in addition to uniform DIF. In Uniform DIF, one group constantly gains an advantage throughout all ability level. Nevertheless, with nonuniform DIF, the conditional dependency shifts and reverses at various points on the ability level rather than providing the reference group with a constant advantage over the ability continuum. Thus, these three DIF detection methods are considered to represent a wide range of statistical procedures. From grades 1 through 12, a total of 336 mathematics items were assessed for both boys and girls in terms of DIF.

In the logistic regression DIF detection method, the total score, grouping variable, and interaction term are included in the model hierarchically. Changes in R^2 values as specified below are interpreted as evidence for uniform DIF or nonuniform DIF (Zumbo, 1999). There are two mainly accepted criteria for DIF detection: Zumbo and Thomas (1997) proposed that ΔR^2 higher than 0.130 indicates moderate DIF and higher than 0.260 indicates large DIF; Jodoin and Gierl (2001) proposed lower values to detect DIF such as ΔR^2 higher than 0.035 indicates moderate DIF and higher than 0.070 indicates large DIF. In the current study, in Model 1, only the total score; in Model 2, the total score and gender; and in Model 3, the total score, gender, and their interaction were used as predictors, and ΔR^2 value of 0.035 was chosen as the threshold for detecting more items. SPSS 27 was used to conduct logistic regression DIF analysis.

The Mantel-Haenszel DIF detection method creates K two-by-two contingency tables, where K represents the number of discrete total score intervals to match group abilities. For each score interval, the expected and observed ratios are calculated and the difference is tested by the chi-square method (Holland & Thayer, 1986). As the chi-square method is highly affected by the large sample size, the MH D-DIF index is proposed for the evaluation where a negative value indicates that an item favors the reference group over the focal group (Holland & Thayer, 1988). Educational Testing Service (ETS) proposed a criterion to flag DIF items: The MH D-DIF index between 1.00 and 1.50 indicates moderate DIF and higher than 1.50 indicates large DIF (Zieky, 1993). The current study used an MH D-DIF index value of 1.00 as the criterion for detecting more items. DIFAS 5.0 (Penfield, 2005) was used for MH DIF detection analysis.

Item Response Theory Likelihood Ratio Test evaluates whether an item has the same IRT item parameters for the reference and focal groups. In the IRT-LR detection procedure, first, all parameters are forced to be equal for both groups (the compact model). Then, parameters are allowed to vary for the studied item (the augmented model), and the difference between these models are compared with a likelihood ratio (LR) test $G^2 = - 2LL_C - (-2LL_A)$ where -2LL denotes the negative two log-likelihood of the compact and augmented models). This difference follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters estimates between the models (Thissen et al., 1993). As the significance of the chi-square test is easily affected by sample size and is difficult to interpret (Stark et al., 2004), the effect size of the b parameter difference could be used to flag DIF items. Steinberg and Thissen (2006) stated that a b parameter difference of 0.50 or larger represents a large effect. As the sample size was very large in the current study, the b parameter differences were estimated, and items were flagged accordingly. IRTLRDIF v2.0b software (Thissen, 2001) was used for the computations.

If an item is detected to have DIF, then expert opinions are crucial to decide whether the item has bias or not. The important technique for determining test item characteristics (such as content, format, context, or language) that may lead to DIF is expert (or judgmental) review, which involves reviewing items by people who are aware about student learning and may have linguistic or cultural experience (Roth et. al, 2013). Thus, in the current study, items flagged as having DIF were evaluated by three experts. Three experts were included in the evaluation committee: a math teacher, a measurement expert, and an academician with expertise in measurement and evaluation. They all had experience on developing and evaluating mathematics items.

Results

Reliability and Unidimensionality of the Instrument

The reliability of the test scores was evaluated using Cronbach's alpha internal consistency coefficient (See Table 2). For each grade level, internal consistency values were ranging from 0.69 to 0.89. These values suggested that the internal consistency of items in each grade level was acceptable. As the study focused on gender differences, these values were computed independently for each gender group. Cronbach's alpha values of gender groups were quite close. For grade levels 3, 4 and 7, Cronbach's Alpha values were relatively higher for boys. This discrepancy suggested that certain items may have had a poorer correlation with girls' overall scores. Whether or not these items create bias was analyzed in the next section.

Evaluation of unidimensionality according to the ratio-of-first-to-second eigenvalues-greater-than-three rule (Slocum-Gori & Zumbo, 2011) and minimum average partial (MAP) showed that for most of the grade levels there is one general factor. MAP is one of the suggested way of deciding number of factor to retain (Zwick & Velicer, 1986) and estimated by the following application (https://afarukkilic.shinyapps.io/Factor_Analysis_For_All_FAFA).

Table 2

Reliability of Test Scores and Unidimensionality

Grade	Cronbach's alpha (All)	Cronbach's alpha (Boys)	Cronbach's alpha (Girls)	First and second eigenvalues	the ratio-of-first-to-second eigenvalues	Minimum Average Partial (MAP)
1	0.69	0.70	0.68	3.107-1.326	2.34	1
2	0.70	0.71	0.69	3.170-1.179	2.69	1
3	0.71	0.73	0.66	3.341-1.288	2.59	1
4	0.77	0.79	0.74	4.042-1.333	3.03	1
5	0.77	0.78	0.74	4.085-1.387	2.95	1
6	0.82	0.82	0.81	4.960-1.330	3.73	1
7	0.75	0.77	0.72	3.933-1.438	2.74	1
8	0.81	0.82	0.80	4.912-1.550	3.17	1
9	0.85	0.85	0.84	5.901-1.723	3.42	2
10	0.85	0.85	0.83	5.753-1.764	3.26	2
11	0.87	0.88	0.84	6.331-1.542	4.11	1
12	0.89	0.89	0.88	7.236-1.432	5.05	1

DIF Results

DIF analyses were conducted using Logistic Regression, Mantel-Haenszel, and Item Response Theory Likelihood Ratio Test DIF detection methods (See Table 3 through Table 8). An item that was flagged by at least two methods was considered as showing DIF. The LR DIF results were reported based on ΔR^2. In terms of both uniform and non-uniform DIF, all ΔR^2 values were lower than 0.035. Thus, it was concluded that the LR DIF method did not detect any item showing DIF in any grade level. The MH results were reported based on the MH D-DIF index. The results revealed that none of the items had an MH D-DIF value of more than 1.00. Thus, according to the MH DIF detection method, none of the items was flagged in any grade level. In the IRT-LR detection procedure, the b parameter differences were reported and the difference of 0.50 or larger was used as DIF detection. For Grade 1 and Grade 9, none of the items; for Grade 2, items 1, 6, 11; for Grade 3, items 1, 2, 3, 11; for Grade 4, items 2, 4, 18; for Grade 5, items 1, 3, 4, 27; for Grade 6, items 1, 4, 5, 27, 29; for Grade 7, item 15; for Grade 8, item 4; for Grade 10, items 1 and 15; for Grade 11, item 11, and for Grade 12, items 3, 4,10 were flagged as having DIF.

Table 3

Grade 1 and Grade 2 DIF Results

	Grade1				Grade2			
Item No	*LR ΔR²* M_2-M_1	*LR ΔR²* M_3-M_2	*ΔMH*	*Δb*	*LR ΔR²* M_2-M_1	*LR ΔR²* M_3-M_2	*ΔMH*	*Δb*
1	0.002	0.000	-0.300	-0.070	0.002	0.000	-0.302	**0.720***
2	0.000	0.003	-0.066	-0.460	0.000	0.000	-0.112	-0.270
3	0.001	0.000	0.096	0.150	0.000	0.001	-0.054	-0.360
4	0.005	0.000	0.297	0.340	0.003	0.000	0.231	0.340
5	0.000	0.000	-0.003	0.040	0.001	0.000	-0.148	-0.050
6	0.000	0.000	0.043	-0.480	0.000	0.003	0.042	**-0.610***
7	0.005	0.001	-0.362	-0.240	0.010	0.000	-0.482	-0.270
8	0.001	0.000	-0.073	0.230	0.001	0.000	-0.111	-0.120
9	0.000	0.001	-0.105	-0.300	0.002	0.000	-0.185	-0.100
10	0.000	0.000	-0.068	-0.180	0.001	0.000	0.154	0.380
11	0.000	0.001	0.023	0.070	0.001	0.001	0.168	**0.500***
12	0.002	0.001	-0.194	-0.300	0.003	0.000	-0.240	-0.200
13	0.002	0.000	-0.181	-0.210	0.003	0.001	-0.254	-0.170
14	0.001	0.000	-0.094	-0.160	0.001	0.000	-0.162	-0.260
15	0.003	0.001	-0.255	0.110	0.001	0.000	-0.113	-0.080
16	0.000	0.000	0.019	0.420	0.000	0.000	-0.041	-0.230
17	0.001	0.000	0.068	-0.060	0.001	0.000	0.104	0.020
18	0.002	0.001	0.197	-0.010	0.001	0.000	0.118	-0.020
19	0.000	0.000	0.085	-0.100	0.001	0.000	-0.166	-0.030
20	0.000	0.001	0.035	0.200	0.000	0.000	0.080	0.000
21	0.000	0.000	0.119	0.000	0.005	0.000	0.323	0.090
22	0.002	0.000	0.177	0.110	0.002	0.000	0.212	0.080
23	0.001	0.000	0.143	-0.040	0.007	0.000	0.373	0.010
24	0.003	0.000	0.247	0.130	0.001	0.000	0.134	0.030

Note: * indicates the item shows DIF

Table 4

Grade 3 and Grade 4 DIF Results

	Grade3				Grade4			
Item No	*LR ΔR²* M_2-M_1	*LR ΔR²* M_3-M_2	*ΔMH*	*Δb*	*LR ΔR²* M_2-M_1	*LR ΔR²* M_3-M_2	*ΔMH*	*Δb*
1	0.003	0.000	0.239	**0.550***	0.004	0.000	0.298	0.420
2	0.026	0.001	0.773	**0.510***	0.033	0.001	0.838	**0.730***
3	0.007	0.001	-0.361	**-0.510***	0.006	0.000	-0.326	-0.200
4	0.003	0.001	-0.192	-0.430	0.003	0.001	-0.222	**-2.140***
5	0.001	0.000	0.066	0.010	0.000	0.000	0.036	0.120
6	0.001	0.000	-0.137	-0.060	0.001	0.000	-0.091	-0.020
7	0.002	0.000	0.268	0.150	0.002	0.000	0.226	0.200
8	0.001	0.000	-0.084	-0.020	0.002	0.000	-0.173	-0.250
9	0.000	0.000	-0.055	0.000	0.000	0.000	-0.002	0.120
10	0.000	0.000	0.056	0.090	0.000	0.001	0.054	0.030
11	0.012	0.001	-0.451	**-0.640***	0.010	0.000	-0.425	-0.360
12	0.001	0.000	-0.128	-0.160	0.002	0.000	-0.217	-0.260
13	0.000	0.000	0.065	-0.420	0.001	0.001	0.124	-0.030
14	0.000	0.001	0.087	0.160	0.000	0.000	-0.065	0.010
15	0.001	0.000	-0.094	-0.160	0.002	0.000	-0.232	-0.160
16	0.000	0.002	0.052	-0.110	0.001	0.000	0.068	0.020
17	0.000	0.000	-0.051	-0.150	0.001	0.000	-0.177	-0.170
18	0.005	0.000	0.319	0.470	0.014	0.000	0.507	**0.730***
19	0.000	0.000	-0.067	-0.150	0.000	0.000	-0.084	-0.060
20	0.001	0.001	0.054	0.060	0.000	0.001	0.029	-0.100
21	0.001	0.000	0.160	-0.020	0.002	0.000	0.187	0.060
22	0.000	0.000	-0.011	-0.170	0.001	0.000	-0.111	-0.010
23	0.001	0.000	-0.081	-0.270	0.000	0.001	-0.012	-0.230
24	0.000	0.000	-0.023	-0.210	0.001	0.000	0.093	0.180

Note: * indicates the item shows DIF

Table 5

Grade 5 and Grade 6 DIF Results

	Grade5				Grade6			
Item No	*LR ΔR²* M_2-M_1	*LR ΔR²* M_3-M_2	*ΔMH*	*Δb*	*LR ΔR²* M_2-M_1	*LR ΔR²* M_3-M_2	*ΔMH*	*Δb*
1	0.008	0.000	-0.470	**-0.620***	0.012	0.000	-0.651	**-0.750***
2	0.003	0.000	0.223	0.170	0.001	0.000	0.099	0.080
3	0.011	0.001	0.464	**0.570***	0.012	0.000	0.483	0.480
4	0.007	0.000	-0.372	**-0.550***	0.005	0.000	-0.349	**-0.550***
5	0.017	0.000	0.650	0.350	0.015	0.000	0.672	**0.580***
6	0.003	0.000	0.277	0.090	0.002	0.000	0.262	0.150
7	0.003	0.006	-0.185	-0.320	0.005	0.004	-0.259	-0.230
8	0.001	0.001	0.058	0.440	0.005	0.002	0.335	0.330
9	0.005	0.000	-0.350	-0.310	0.002	0.001	-0.212	-0.190
10	0.001	0.001	-0.149	-0.030	0.000	0.000	0.054	-0.020
11	0.000	0.001	-0.070	-0.190	0.001	0.000	-0.139	-0.060
12	0.000	0.000	-0.065	-0.230	0.001	0.000	-0.075	-0.360
13	0.003	0.000	-0.240	-0.290	0.002	0.000	-0.201	-0.160
14	0.006	0.000	0.361	0.280	0.011	0.000	0.513	0.190
15	0.000	0.000	-0.059	-0.110	0.002	0.001	-0.193	-0.110
16	0.000	0.000	-0.006	-0.090	0.000	0.000	-0.040	-0.040
17	0.002	0.000	0.158	0.210	0.001	0.000	0.137	0.150
18	0.013	0.000	-0.499	-0.490	0.015	0.000	-0.517	-0.350
19	0.001	0.000	0.123	0.070	0.001	0.001	0.154	0.040
20	0.001	0.000	-0.104	-0.100	0.000	0.000	-0.051	-0.040
21	0.000	0.000	0.076	0.020	0.000	0.000	0.059	-0.160
22	0.001	0.000	-0.093	-0.360	0.001	0.001	-0.191	-0.360
23	0.001	0.000	-0.115	-0.150	0.001	0.000	-0.147	-0.190
24	0.003	0.001	0.263	0.200	0.002	0.000	0.196	0.130
25	0.001	0.000	-0.120	-0.140	0.002	0.000	-0.213	-0.250
26	0.001	0.000	-0.092	0.040	0.000	0.001	-0.039	0.030
27	0.001	0.000	-0.113	**-1.100***	0.004	0.000	-0.241	**-1.200***
28	0.000	0.001	0.146	0.270	0.000	0.001	0.124	0.270
29	0.001	0.001	0.201	0.000	0.002	0.001	0.225	**1.050***
30	0.000	0.000	0.014	0.070	0.002	0.000	0.202	-0.050

Note: * indicates the item shows DIF

Table 6

Grade 7 and Grade 8 DIF Results

	Grade7				Grade8			
Item No	*LR ΔR²* M_2-M_1	*LR ΔR²* M_3-M_2	*ΔMH*	*Δb*	*LR ΔR²* M_2-M_1	*LR ΔR²* M_3-M_2	*ΔMH*	*Δb*
1	0.000	0.001	0.087	0.170	0.002	0.000	0.180	.250
2	0.002	0.001	-0.226	-0.210	0.003	0.004	-0.270	-.180
3	0.005	0.000	0.330	0.310	0.003	0.000	0.288	.240
4	0.000	0.000	-0.042	-0.300	0.000	0.004	0.082	**1.180***
5	0.003	0.002	-0.282	-0.300	0.002	0.003	-0.247	-.180
6	0.000	0.000	-0.014	0.160	0.000	0.001	-0.011	0.340
7	0.001	0.000	-0.152	-0.230	0.000	0.001	-0.061	-0.270
8	0.006	0.001	-0.327	-0.310	0.000	0.000	-0.076	-0.040
9	0.005	0.001	-0.332	0.000	0.005	0.001	-0.306	-0.120
10	0.000	0.000	0.056	0.000	0.000	0.001	-0.058	-0.070
11	0.001	0.000	0.134	0.170	0.001	0.000	0.150	0.150
12	0.000	0.000	0.023	-0.040	0.000	0.000	-0.039	-0.030
13	0.001	0.000	0.156	0.090	0.001	0.000	0.115	0.040
14	0.001	0.000	0.128	-0.080	0.000	0.000	-0.040	-0.220
15	0.022	0.003	0.699	**0.780***	0.031	0.000	0.874	0.470
16	0.000	0.000	0.065	0.030	0.000	0.000	0.013	0.100
17	0.005	0.000	-0.343	-0.200	0.004	0.000	-0.292	-0.230
18	0.000	0.001	-0.044	-0.100	0.001	0.001	0.200	-0.150
19	0.000	0.000	0.109	-0.190	0.001	0.002	0.170	-0.060
20	0.005	0.000	0.346	0.050	0.001	0.000	0.137	0.040

	Grade7				Grade8			
Item No	*LR* ΔR^2 M_2-M_1	*LR* ΔR^2 M_3-M_2	*ΔMH*	*Δb*	*LR* ΔR^2 M_2-M_1	*LR* ΔR^2 M_3-M_2	*ΔMH*	*Δb*
21	0.000	0.001	0.057	-0.110	0.000	0.000	-0.010	0.020
22	0.001	0.000	-0.145	-0.210	0.003	0.000	-0.221	-0.160
23	0.000	0.000	0.034	0.020	0.001	0.000	-0.076	0.100
24	0.005	0.003	0.360	0.120	0.004	0.006	0.337	0.310
25	0.001	0.000	0.094	-0.320	0.000	0.000	-0.016	-0.190
26	0.001	0.001	-0.158	-0.130	0.000	0.000	-0.058	0.050
27	0.002	0.001	-0.207	-0.240	0.002	0.001	-0.218	-0.100
28	0.000	0.001	-0.031	0.080	0.000	0.000	-0.010	0.040
29	0.001	0.000	-0.071	-0.400	0.002	0.000	-0.200	-0.280
30	0.001	0.001	0.199	0.400	0.004	0.000	0.292	0.060

Note: * indicates the item shows DIF

Table 7

Grade 9 and Grade 10 DIF Results

	Grade9				Grade10			
Item No	*LR* ΔR^2 M_2-M_1	*LR* ΔR^2 M_3-M_2	*ΔMH*	*Δb*	*LR* ΔR^2 M_2-M_1	*LR* ΔR^2 M_3-M_2	*ΔMH*	*Δb*
1	0.007	0.002	-0.438	0.250	0.010	0.001	-0.506	**1.280***
2	0.000	0.001	-0.130	-0.110	0.001	0.000	-0.185	-0.060
3	0.000	0.000	-0.060	0.080	0.003	0.000	-0.245	-0.020
4	0.000	0.001	-0.004	-0.050	0.000	0.001	-0.036	0.130
5	0.000	0.000	-0.031	-0.010	0.000	0.001	-0.025	-0.090
6	0.000	0.001	0.112	0.010	0.000	0.000	-0.081	-0.050
7	0.001	0.000	-0.159	-0.020	0.002	0.000	-0.192	-0.260
8	0.005	0.000	-0.399	-0.270	0.006	0.000	-0.402	-0.210
9	0.002	0.000	0.253	0.110	0.002	0.000	0.227	0.300
10	0.000	0.001	0.167	0.030	0.000	0.001	0.083	-0.060
11	0.000	0.000	-0.044	-0.050	0.000	0.000	-0.059	0.010
12	0.001	0.001	0.150	-0.200	0.002	0.000	0.194	0.190
13	0.000	0.000	-0.109	-0.080	0.000	0.000	-0.059	-0.090
14	0.000	0.000	0.052	0.030	0.000	0.000	-0.041	0.050
15	0.000	0.002	-0.065	-0.460	0.001	0.004	-0.137	**0.670***
16	0.001	0.000	-0.162	-0.160	0.001	0.001	-0.173	-0.150
17	0.016	0.001	0.708	0.440	0.007	0.000	0.433	0.290
18	0.000	0.001	0.128	0.100	0.001	0.000	-0.145	-0.160
19	0.000	0.003	-0.113	0.270	0.000	0.000	0.076	-0.030
20	NA	NA	NA	NA	NA	NA	NA	NA
21	0.001	0.000	0.118	0.080	0.000	0.000	0.088	-0.070
22	0.002	0.005	0.234	-0.120	0.012	0.001	0.633	0.020
23	0.001	0.000	0.089	0.060	0.005	0.000	0.372	0.190
24	0.001	0.001	0.175	-0.190	0.001	0.000	0.147	0.160
25	0.000	0.000	-0.122	-0.070	0.000	0.000	-0.043	-0.150
26	0.000	0.000	0.012	0.010	0.000	0.000	-0.034	-0.140
27	0.000	0.000	-0.043	-0.040	0.000	0.000	-0.040	-0.090
28	0.004	0.001	-0.367	-0.210	0.003	0.001	-0.284	-0.270
29	0.007	0.003	-0.435	-0.410	0.002	0.001	-0.211	-0.170
30	0.005	0.003	0.385	-0.400	0.010	0.000	0.520	0.140

Note: * indicates the item shows DIF; NA: The item was canceled due to printing issues.

Table 8

Grade 11 and Grade 12 DIF Results

	Grade11				Grade12			
Item No	*LR* ΔR^2 M_2-M_1	*LR* ΔR^2 M_3-M_2	*ΔMH*	*Δb*	*LR* ΔR^2 M_2-M_1	*LR* ΔR^2 M_3-M_2	*ΔMH*	*Δb*
1	0.000	0.001	0.047	0.240	0.000	0.000	-0.025	-0.060
2	0.000	0.000	-0.133	-0.020	0.007	0.000	-0.527	-0.160
3	0.017	0.006	-0.598	0.140	0.004	0.002	0.873	**1.070***
4	0.006	0.001	-0.355	-0.210	0.003	0.004	0.291	**0.750***
5	NA	NA	NA	NA	NA	NA	NA	NA

	Grade11				Grade12			
Item No	*LR ΔR²* M_2-M_1	*LR ΔR²* M_3-M_2	*ΔMH*	*Δb*	*LR ΔR²* M_2-M_1	*LR ΔR²* M_3-M_2	*ΔMH*	*Δb*
6	0.003	0.000	-0.199	-0.240	0.001	0.002	0.137	0.150
7	0.000	0.002	-0.001	0.200	0.001	0.002	0.000	0.240
8	0.001	0.001	0.178	0.320	0.004	0.000	0.277	0.400
9	0.005	0.001	-0.388	-0.190	0.005	0.001	-0.568	-0.260
10	0.004	0.003	-0.294	0.030	0.006	0.008	-0.346	**-0.540***
11	0.026	0.000	0.774	**0.970***	0.006	0.000	0.237	0.470
12	0.004	0.000	-0.342	-0.240	0.000	0.015	0.030	-0.250
13	0.002	0.002	-0.246	-0.180	0.010	0.005	-0.684	-0.080
14	0.001	0.000	0.125	0.270	0.010	0.000	0.449	0.390
15	0.009	0.001	0.413	0.410	0.023	0.000	0.708	0.330
16	0.018	0.001	0.659	0.210	0.014	0.001	0.545	0.140
17	0.000	0.001	0.086	-0.030	0.004	0.004	0.409	0.380
18	0.010	0.000	-0.557	-0.330	0.009	0.006	-0.602	-0.150
19	0.001	0.000	0.097	-0.050	0.001	0.001	-0.283	-0.030
20	0.003	0.001	-0.275	-0.080	0.008	0.000	-0.608	-0.200
21	0.001	0.000	-0.141	-0.070	0.000	0.002	-0.042	0.050
22	0.000	0.001	0.107	-0.140	0.014	0.001	0.430	0.080
23	0.000	0.000	0.127	0.230	0.003	0.000	-0.323	-0.020
24	0.000	0.000	-0.166	-0.060	0.003	0.001	0.283	0.270
25	0.001	0.000	0.094	-0.180	0.006	0.002	-0.513	-0.040
26	0.001	0.001	0.184	0.190	0.001	0.001	-0.129	-0.120
27	0.000	0.001	-0.023	0.090	0.001	0.000	0.192	0.090
28	0.000	0.001	-0.072	-0.010	0.001	0.009	-0.135	0.020
29	NA	NA	NA	NA	NA	NA	NA	NA
30	0.000	0.000	-0.016	-0.020	0.000	0.003	-0.119	-0.300

Note: * indicates the item shows DIF; NA: The item was canceled due to printing issues.

Evaluation of DIF Results

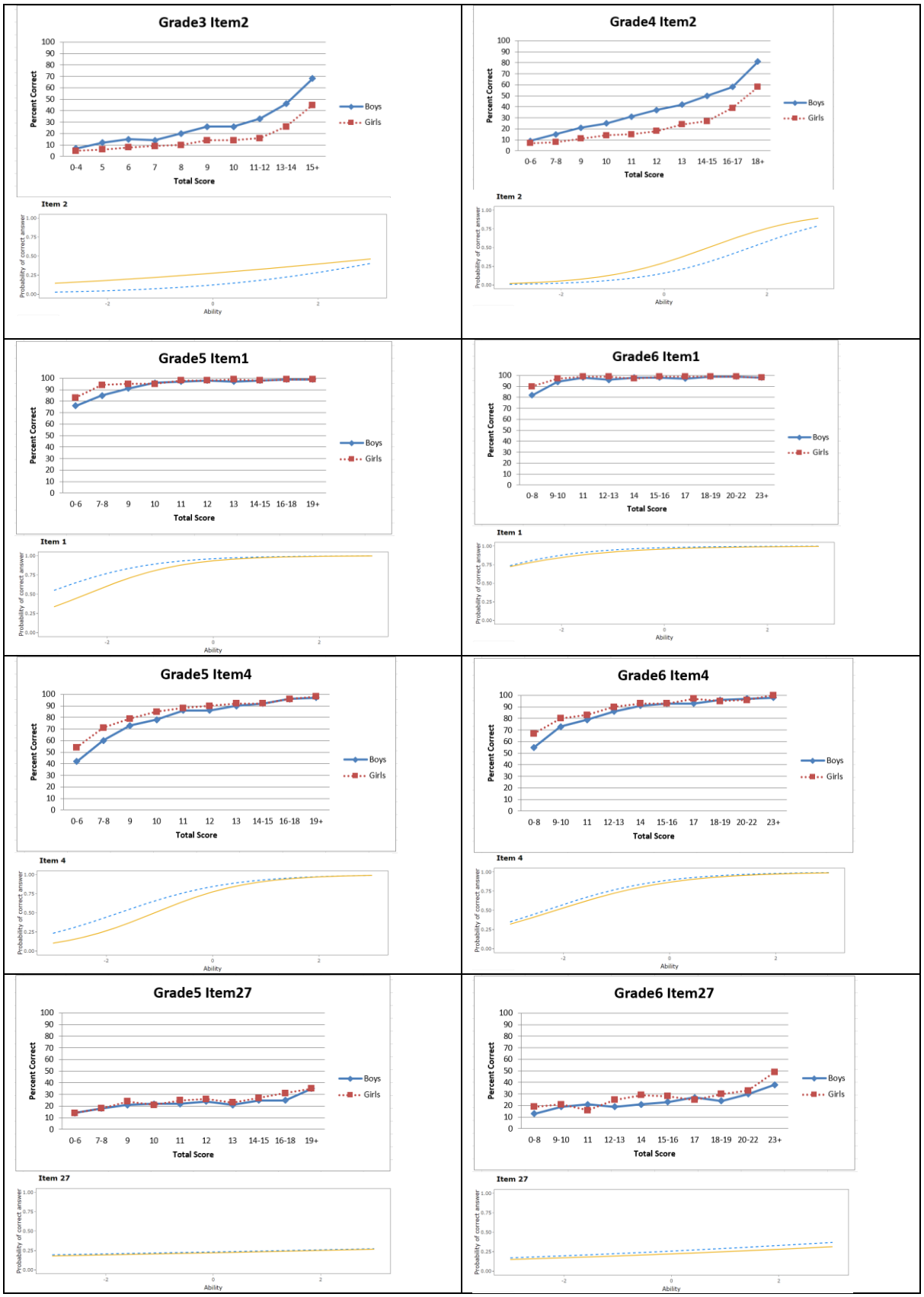
Overall, out of 336 items, LR DIF and MH flagged none of the items whereas IRT-LR flagged 27 items as having DIF with 16 items favoring boys, and 11 items favoring girls. As none of the items was flagged by two or more DIF detection methods, it was concluded that 2022 Kangaroo Mathematics items were free of uniform or nonuniform DIF and bias across the gender groups.

However, an item flagged as showing DIF in both grade levels (as each item was administered to two grade levels) was worth examining. Thus, items 2 in Grades 3 and 4; and items 1, 4 and 27 in Grades 5 and 6 were evaluated by the experts to determine if these items contain any bias (these four items are presented in the appendix). Additionally, to provide more information for this examination, percentages of correct responses and ICC's for gender groups were drawn (see Figure 2). In percentage correct graphs, the x-axis represents the total score groups created according to deciles, and the y-axis represents the percentage correct responses of gender groups. In ICCs x-axis represents the ability score estimated by IRT, and the y-axis represents the probability of correct answer of gender groups. These graphs indicated that, controlling student ability, the curves did not differ seriously.

Following the process of detecting DIF items empirically, there is a post hoc procedure to examine DIF items judgmentally to detect the possible causes of DIF, if any. First, the direction of DIF was determined. Item 2 in Grades 3 and 4 favored boys; whereas, items 1, 4 and 27 in Grades 5 and 6 favored girls. Then, these four items were evaluated by three experts. The experts were given the items and asked if they anticipated any DIF. The DIF detection results such as the direction of DIF items were also presented. Possible sources of DIF were highlighted to make experts to focus on those sources. The experts claimed that the item wording or pictures of the items were free of gender bias. Thus, all experts concluded that these four items did not contain any gender-specific content to create gender bias.

Figure 2

Percentages of correct responses and ICC's for gender groups



Evaluation of Gender Mean Score Differences Excluding DIF Items

As it was shown in Table 1, boys perform better than girls; however, these differences were small according to Cohen's d (Cohen, 1988), except for grade 11. Effect size and the standardized mean difference allow for comparing the difference between groups without being affected by sample size (Field, 2013). In this part, to evaluate the negative consequences associated with the presence of DIF, the original effect sizes and the effect sizes excluding DIF items were compared (see Table 9). Overall, the effect sizes of the gender differences were very similar with or without DIF items. Thus, it was concluded that these DIF items did not produce any biased consequences on scores.

Table 9. Effect Sizes of with and without DIF items

Grade	Effect Size (original)	Effect Size (excluding DIF)
Grade 1	0.14*	0.14*
Grade2	0.09*	0.08*
Grade3	0.15*	0.15*
Grade4	0.18*	0.13*
Grade5	0.30*	0.29*
Grade6	0.28*	0.28*
Grade7	0.22*	0.20*
Grade8	0.28*	0.28*
Grade9	0.32*	0.32*
Grade10	0.35*	0.36*
Grade11	0.43*	0.41*
Grade12	0.18	0.18

* $p < 0.05$

Discussion

Mathematical problem-solving competitions provide opportunities for mathematicians and mathematics educators to collaborate, create tests together and conduct joint research (de Losada & Taylor, 2022). It is considered that scholars from the field of educational measurement will involve in these tests in the near future; and by looking through the lenses of measurement specialists, the research on the psychometric properties of these competition tests will increase. Focusing on a well-known mathematical problem-solving competition, the research papers on Kangaroo Mathematics competitions are mainly related to the evaluation of content (Jiang, & Xiong, 2021), description of the competition, and problems (Akveld, Caceres-Duque, & Geretschläger, 2020; Akveld, Caceres-Duque, Nieto Said & Sánchez Lamoneda, 2020), test-wiseness strategies (Donner et al., 2021), gender gap and task relationship (Applebaum & Leikin, 2019), the effect of teacher gender (Escardibul & Mora, 2013), and comparing student performances in competitions and classroom tests (Mellroth, 2015). There is a lack of research on the psychometric properties of mathematical problem-solving competitions. As a result, the current study addressed the issue of gender equity through the lenses of psychometry and examined Kangaroo mathematics competition items for DIF and bias.

The preliminary results showed that boys had higher mathematics scores in the competition than girls, and these differences were mainly small according to Cohen's d (Cohen, 1988). When there is a difference between two groups, the difference could be due to either an impact or bias. To understand the nature of the differences, DIF analyses were conducted. Out of 336 items, none of the items displayed DIF on two DIF detection methods (LR DIF and MH) and 27 items displayed DIF on the IRT-LR method, 16 items favoring boys and 11 items favoring girls. Thus, it was concluded that there was not an issue of DIF across gender groups because no items were detected by two or more DIF detection methods. To have a deeper investigation, items that displayed DIF in one method in both grade levels were identified. These four items were examined based on the direction of DIF (three out of four items favored girls), percentages of correct responses for gender groups, and expert evaluations. As a result of this examination, no evidence for bias was found. Finally, the effect sizes with or without DIF items (omitting 27 items flagged by IRT-LR) were compared and no change in effect sizes was observed. Overall, based all of these detailed investigations, it was concluded that there was no gender bias on Kangaroo Mathematics competition items. Thus, the small sized difference observed between gender groups could not be explained by bias.

Examining DIF is not only related to individual test scores, but also to the educational quality of the assessment instruments by providing validity evidence (Berrío et al., 2020). The standards for educational and psychological testing require test publishers to report the evidence of reliability and validity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Evaluating the possible bias and providing evidence for bias-free assessment scores are essential for valid score interpretations. Increasing fairness in testing increases validity in scores and helps minimize the construct-irrelevant variances (ETS, 2022). Thus, like other large-scale assessments, mathematical competitions are also needed to provide fair and valid assessment across subgroups. By evaluating bias in one of the major mathematical competitions, the current study is expected to lead to other studies that focus on the psychometric properties of mathematics competitions.

It is vital to highlight that fairness begins with the preparation of the test plan and the development of the items, because introducing unfair content or construct-irrelevant variance for some groups of examinees may result in bias (ETS, 2022). For instance, considering gender bias, the inclusion of items based on soccer-related calculations on the league table could be a source of bias. Similarly, having items related to metro systems of metropolitan cities may create bias across students located in rural and urban regions. As validity is related to the interpretation and consequences of test scores (Messick, 1989), an item that could have a detrimental consequence on a certain group needs to be avoided. Thus, test developers and test reviewers need to be cautious about the issue of bias. Additionally, test takers need to be knowledgeable about this issue and demand bias-free assessments.

The current study concentrated on a gender bias in data from a single country. It is suggested that future studies compare international mathematical competitions cross-culturally. Evaluating the reliability and validity of these competitions across cultures and evaluating measurement invariance at the test-level and DIF at the item-level would add valuable information to the literature. International large-scale assessments such as PISA and TIMSS have provided an assessment framework before the administration and technical reports after the administration. Similarly, reporting psychometric findings of international mathematics competitions on an annual basis as technical reports would also increase the interest of more scholars and stakeholders throughout the world. In addition to gender groups and country-level comparisons, there are other groups to consider for the fairness of test results. ETS (2022) recommended that special attention should be paid to the following groups on any assessment: age, appearance, citizenship status, disability, ethnicity, gender, national or regional origin, native language, race, religion, sexual orientation and socioeconomic status. Thus, future studies could examine DIF across these groups in mathematics competitions.

This mathematics competition was also given to first graders. First graders are still developing their reading skills. Thus, even though they volunteered to participate in this competition, their score might have construct-irrelevant variance. Given that the first graders' scores had an adequate Cronbach's alpha, it appears that there was no issue regarding their age. Nonetheless, studying first graders' experiences in these competitions might yield important data for subsequent studies.

References

Akveld, M., Caceres-Duque, L. F., & Geretschläger, R. (2020). Math Kangaroo. *Mathematics Competitions*, *33*(2), 48–66.

Akveld, M., Caceres-Duque, L. F., Nieto Said, J. H., & Sánchez Lamoneda, R. (2020). The Math Kangaroo Competition. *Espacio Matemático*, *1*(2), 74–91.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, *36*(3), 185–198. https://doi.org/10.1111/j.1745-3984.1999.tb00553.x

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

Andritsch, L., Hauke, E., & Kelz, J. (2020). How to create and solve: Analysis of items from the Mathematical Kangaroo from two perspectives. In R. Geretschläger (Ed.), *Engaging young students in mathematics through competitions—World perspectives and practices, Vol. II: Mathematics competitions and how they relate to research, teaching and motivation* (pp. 117–136). World Scientific.

Applebaum, M., & Leikin, R. (2019). Girls' performance in the Kangaroo contest. In M. Nolte (Ed.), *Including the Highly Gifted and Creative Students–Current Ideas and Future Directions-Proceedings of the 11th International Conference on Mathematical Creativity and Giftedness (mcg 11)*, (pp. 87–94). Hamburg, Germany.

Arikan, S. (2019). Are Differentially Functioning Mathematics Items Reason of Low Achievement of Turkish Students in PISA 2015? Journal of Measurement and Evaluation in Education and Psychology, 10(1), 49–67. https://doi:10.21031/epod.466860

Baniasadi, A., Salehi, K., Khodaie, E., Bagheri Noaparast, K., & Izanloo, B. (2023). Fairness in classroom assessment: A systematic review. *The Asia-Pacific Education Researcher*, *32*, 91–109. https://doi.org/10.1007/s40299-021-00636-z

Berrío, Á. I., Gomez-Benito, J., & Arias-Patiño, E. M. (2020). Developments and trends in research on methods of detecting differential item functioning. *Educational Research Review*, *31*, 100340. https://doi.org/10.1016/j.edurev.2020.100340

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, (2nd ed.). Hillsdale, NJ: Erlbaum.

de Losada, M. F., & Taylor, P. J. (2022). Perspectives on mathematics competitions and their relationship with mathematics education. *ZDM–Mathematics Education*, *54*(5), 941–959. https://doi.org/10.1007/s11858-022-01404-z

Desjarlais, M. A. (2009). *Gender differences on the American Mathematics Competition AMC 8 contest*. The University of Nebraska-Lincoln.

Donner, L., Kelz, J., Stipsits, E., & Stuhlpfarrer, D. (2021). Which test-wiseness based strategies are used by Austrian winners of the Mathematical Kangaroo?. *Mathematics Competitions*, *34*(1), 88–101.

Dorans, N. J. (2013). ETS contributions to the quantitative assessment of item, test, and score fairness. *ETS Research Report Series*, *2013*(2), i–38.

Escardibul, J. O., & Mora, T. (2013). Teacher gender and student performance in Mathematics. Evidence from Catalonia (Spain). *Journal of Education and Training Studies*, *1*(1), 39–46.

ETS. (2022). ETS guidelines for developing fair tests and communications. https://www.ets.org/content/dam/ets-org/pdfs/about/fair-tests-and-communications.pdf

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London: Sage.

Geretschläger, R., & Donner, L. (2022). Writing and choosing problems for a popular high school mathematics competition. *ZDM–Mathematics Education*, *54*(5), 971–982. https://doi.org/10.1007/s11858-022-01351-9

Gneezy, U., Muriel, N., & Aldo, R. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics*, *118*(3), 1049–1074. https://doi.org/10.1162/00335530360698496

He, J., & van de Vijver, F. J. R. (2013). Methodological issues in cross-cultural studies in educational psychology. In G. A. D. Liem & A. B. I. Bernardo (Eds.), *Advancing cross-cultural perspectives on educational psychology: A festschrift for Dennis McInerney (*pp. 39–56). Charlotte, NC: Information Age Publishing.

Holland, P. W., & Thayer, D. T. (1986). *Differential item functioning and the Mantel-Haenszel procedure* (ETS Research Report No. RR-86-31). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity*, (pp.129–145). Hillsdale, N.J.: Erlbaum.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, *321*(5888), 494–495. https://doi.org/10.1126/science.1160364

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing*, *1*(2), 93–114. https://doi.org/10.1207/S15327574IJT0102_1

Jiang, P., & Xiong, B. (2021, April). Analyze the quality of Math Kangaroo problems with a content analysis. In *Journal of Physics: Conference Series* (Vol. 1875, No. 1, p. 012015). IOP Publishing.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2

Kankaraš, M., & Moors, G. (2014). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology*, *45*(3), 381–399. https://doi.org/10.1177/0022022113511297

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

Mellroth, E. (2015). Problem solving competency and the mathematical kangaroo. In K. Krainer & N. Vondrová (Eds.), *Proceedings of the Ninth Congress of the European Society for Research in Mathematics Education (CERME9, 4-8 February 2015)* (pp. 1095–1096). Prague, Czech Republic: Charles University in Prague, Faculty of Education and ERME. *https://hal.science/CERME9/public/CERME9_Proceedings_2015.pdf*

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd edition, pp.13–103). New York, NY: Macmillan.

Niederle, M., & Vesterlund, L. (2010). Explaining the gender gap in Math test scores: The Role of Competition. *Journal of Economic Perspectives*, *24*(2), 129–144. https://doi.org/10.1257/jep.24.2.129

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement*, *29*, 150–151. https://doi.org/10.1177/0146621603260686

Reynolds, K., Khorramdel, L., & von Davier, M. (2022). Can students' attitudes towards mathematics and science be compared across countries? Evidence from measurement invariance modeling in TIMSS 2019. *Studies in Educational Evaluation*, *74*, 101169. https://doi.org/10.1016/j.stueduc.2022.101169

Roberson, N. D., & Zumbo, B. D. (2019). Migration background in PISA's measure of social belonging: Using a diffractive lens to interpret multi-method DIF studies. *International Journal of Testing*, *19*(4), 363–389. https://doi.org/10.1080/15305058.2019.1632316

Roth, W.-M., Oliveri, M. E., Sandilands, D., Lyons-Thomas, J., & Ercikan, K. (2013). Investigating sources of differential item functioning using expert think-aloud protocols. *International Journal of Science Education*, *35*, 546–576. https://doi.org/10.1080/09500693.2012.721572

Rubright, J. D., Jodoin, M., Woodward, S., & Barone, M. A. (2022). Differential item functioning analysis of United States medical licensing examination step 1 items. *Academic Medicine*, *97*(5), 718–722. https://doi.org/10.1097/ACM.0000000000004567

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research*, *102*, 443–461. https://doi.org/10.1007/s11205-010-9682-8

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important?. *Journal of Applied Psychology*, *89*(3), 497–508. https://doi.org/10.1037/0021-9010.89.3.497

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological methods*, *11*(4), 402–415. https://doi.org/10.1037/1082-989X.11.4.402

Thissen, D. (2001). IRTLRDIF v2.0b: *Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning* [Documentation for computer program]. L.L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill.

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

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis of comparative research*. Thousand Oaks, CA: Sage.

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, *54*(2), 119–135. https://doi.org/10.1016/j.erap.2003.12.004

Wedman, J. (2018). Reasons for gender-related differential item functioning in a college admissions test. *Scandinavian Journal of Educational Research*, *62*(6), 959–970. https://doi.org/10.1080/00313831.2017.1402365

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning, (*337–364). Hillsdale, NJ: Lawrence Erlbaum.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832

Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research & Policy Studies*, *5*(1), 1–23.

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Prince George, Canada: Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP History assessment. *Journal of Educational Measurement, 26*, 55–66. https://doi.org/10.1111/j.1745-3984.1989.tb00318.x

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432–442. https://doi.org/10.1037/0033-2909.99.3.432

Uluslararası Matematik Yarışması Maddelerinin Cinsiyet Gruplarına Göre Değişen Madde Fonksiyonu (DMF) açısından İncelenmesi

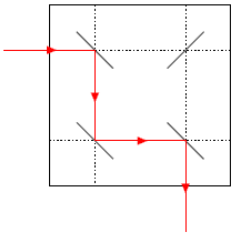
Öz

Matematiksel problem çözme yarışmaları bir asırdan fazla bir süredir uygulanmaktadır. Araştırmalar bu yarışmalarda kızlar ve erkekler arasında başarı farkı olduğunu göstermektedir. Cinsiyet grupları arasındaki puan farkının gerçekte var olan bir farklılıktan mı yoksa sınavdaki sorulardan mı kaynaklandığının belirlenmesi değerli bilgiler sunacaktır. Bu nedenle bu çalışma pek çok öğrencinin katıldığı matematik yarışmalarından biri olan Uluslararası Kanguru Matematik yarışmasındaki madde yanlılığını incelemektedir. Maddelerin Değişen Madde Fonksiyonu (DMF) içerme durumlarını incelemek için Lojistik Regresyon, Mantel-Haenszel ve Madde Tepki Kuramı Olabilirlik Oranı Testi kullanılmıştır. Analizlerden elde edilen bulgulara göre toplam 336 maddeden oluşan bu matematik testlerinin DMF içermediği, dolayısıyla maddelerin cinsiyet grupları arasında yanlılık yaratmadığı sonucuna varılmıştır. Makalede geçerlilik ve yanlılıkla ilgili diğer çıkarımlar ayrıntılı olarak tartışılmıştır.

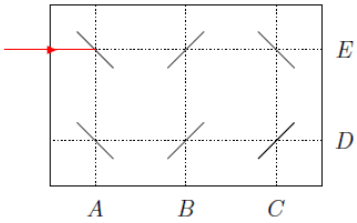
Anahtar kelimeler: DMF, yanlılık, uluslararası matematik yarışmaları, lojistik regresyon, Mantel-Haenszel ve Madde Tepki Kuramı Olabilirlik Oranı Testi

Appendix A

Grade 3-4, Item 2



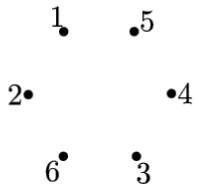
2. The mirrors reflect the laser beam like this:
Where does this laser beam end?



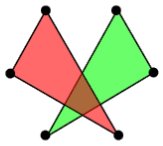
(**A**) A (**B**) B (**C**) C (**D**) D (**E**) E

Grade 5-6, Item 1

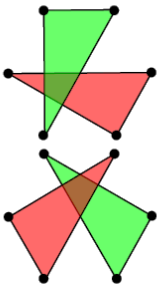
There are six points numbered as shown. We create two triangles, one by connecting the points with even numbers, the other one by connecting the points with odd numbers. Which of the five figures do we get?



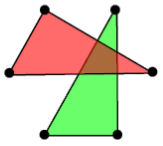
(**A**)



(**B**)

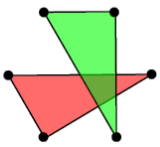


(**C**)



(**D**)

(**E**)



Grade 5-6, Item 4

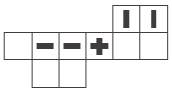
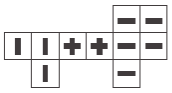
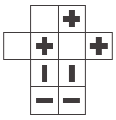
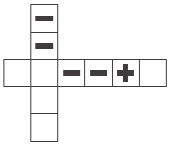
Cistercian numerals were used in the early thirteenth century. Any integer from 1 to 99 can be represented by a single glyph, combining the glyphs below.

So 24 looks like  , 81 looks like  and 93 looks like  . How does 45 look like?

(A)  (B)  (C)  (D)  (E) 

Grade 5-6, Item 27

Which of the following nets cannot be folded into the solid  ?

(A)  (B)  (C)  (D)  (E)