Title: Sex differences in 32,347 Jordanian 4th graders on the national exam of mathematics

Authors: Ismael S. Al-Bursan¹, Emil O. W. Kirkegaard², John Fuerst³, Salaheldin F. Bakhiet⁴, Mohammad F. Al. Qudah⁵, Adel S. Abduljabbar⁶

Supplementary material and acknowledgments

(This information cannot easily be blinded. However, the R output can be included in the supplementary materials in blinded version as HTML.)

The R notebook can be found at http://rpubs.com/EmilOWK/Jordan_4th_graders_sex (as well as in the supplementary materials).

Supplementary materials including code, high quality figures and data can be found at <https://osf.io/gu5y7/files/>.

Conflicts of interests

Conflicts of interest: none.

Suggested reviewers

Jelte Wicherts. Familiar with IRT-modeling, measurement invariance and group differences.

Richard Lynn. Extensive familiarity with the sex differences literature. The finding is related to his model.

Paul Irwing. Familiar with the sex differences literature and complex modeling thereof.

Roberto Colom. Familiar with item-level analyses of sex differences.

All four suggestions are authors whose published work is cited in this study.

1 King Saud university, college of education, department of psychology, email: ibursan@ksu.edu.sa

2 Corresponding author. Ulster Institute for Social Research, United Kingdom. emil@emilkirkegaard.dk.

3 Ulster Institute for Social Research, United Kingdom. j122177@hotmail.com

4 King Saud university, college of education, department of special education, email: slh9999@yahoo.com

5 King Saud university, college of education, department of psychology, email: malqudah@ksu.edu.sa

6 King Saud university, college of education, department of psychology, email: abduljabbar@ksu.edu.sa

Sex differences in 32,347 Jordanian 4th graders on the national exam of mathematics

Abstract

Sex differences in mathematical ability were examined in a nation-wide sample of 32,346 Jordanian 4th graders (age 9-10) on a 40-item mathematics test. Overall, girls were found to perform slightly better ($d = -0.12$), while boys had slightly more variation in scores (standard deviations of 1.02 and 0.98 for boys and girls, respectively). 2 parameter item response theory analysis showed that the item parameters were almost identical across sexes; correlations were .97 [95% confidence interval: .95 to .99] and .90 [.82 to .95] for difficulties and loadings, respectively. A bootstrapping analysis revealed that these could not be explained by sampling error alone, but reflected genuine departures from measurement invariance. However, sex differences in items showed no relationship to the relative size of the departure from measurement invariance. A correlation between item difficulty and male advantage was observed, $r = .57$ [.32 to .75] (.68 after exclusion of an outlier), such that the relative male advantage increased with increased item difficulty. This finding was essentially unchanged in multiple regression with other item characteristics included as controls. Results were interpreted to be congruent with Richard Lynn's developmental theory of sex differences in intelligence as well as other models.

Key words: sex differences, intelligence, mathematics, cognitive ability, item data, IRT, item response theory, Jordan

Highlights:

- 32,346 Jordanian 4th graders took a 40-item test of mathematical ability.
- The test showed a slight departure from measurement invariance.
- Overall, girls slightly outperformed boys ($d = -0.12$).
- Boys varied slightly more in their scores (ratio of sd's = 1.04).
- Boys performed relatively better on the more difficult items ($r = 0.57$).

1. Introduction

Sex differences in cognitive abilities have been widely studied and discussed in the last few decades (Halpern, 2013; Halpern et al., 2007; Irwing & Lynn, 2005). Previous research on mathematical abilities have tended to find either no differences or small male advantages (Whitley et al., 2016), depending on the nature of the test and the age of the sample. Most previous studies were based on (sub)test level data, not item level. However, there is an increasing tendency towards analyzing item data instead, as this provides more information about sources of observed differences including measurement bias. Furthermore, the studies found in the literature typically concern Western samples (Henrich, Heine, & Norenzayan, 2010), while other regions of the world are rarely studied. One rarely studied country is Jordan, a small Middle Eastern country located on the Arab peninsula and which borders Israel, Syria, Iraq and Saudi Arabia. The purpose of this study was to examine possible sex differences and measurement issues in mathematical ability in Jordan using item data so as to maximize the informativeness of the findings.

2. Test design and data collection

The test was based on the TIMSS 2015 test and was prepared at the Jordanian Ministry of Education, Directorate of tests, by a committee composed of three educational supervisors of mathematics, an educational supervisor of measurement and evaluation, and a mathematics teacher for 4th grade. Content validity was verified by a committee composed of 3 mathematics teachers, 3 math educational supervisors, 2 of specialists in measurement and evaluation, and 2 of university professors in the mathematics curriculum. The first version of the test contained 48 items, and was then reduced to 43 as a result of a content revision. Before the main study, a pilot study was done to screen for problematic items. This was done on a sample of 157 pupils (83 male; 53%). Based on this, 3 items were removed due to low discrimination (< 0.20).

After this, the remaining 40 items were administered to 4th graders (age 9-10) in all schools in Jordan in the academic year of 2014/2015, with 50% of students in each school participating. The resulting sample size was 32,346 (16,400 male; 51%).

3. Analyses

All analyses were done in R (3.4.0; R Core Team, 2017). An R notebook of the code and output is available in the supplementary materials.

3.1. Main IRT analysis

The item data were analyzed with item response theory (IRT; DeMars, 2010). IRT analysis was done using the **psych** package (Revelle, 2017) using the 2 parameter normal (2PN) model. This package also performs factor analysis on the latent correlation matrix (tetrachoric; see Uebersax, 2015). This allows the calculation of item loadings that are on the same scale as regular test-level loadings, and which are

not affected by item pass rates unlike item-whole Pearson correlations (Wicherts, n.d.; Wicherts & Johnson, 2009). Finally, we calculated sex gaps in d scores by converting pass rates into z scores based on the normal distribution. As before, these are not affected by item pass rate differences, unlike difference in pass rates.⁷

Table 1 shows the item parameters.

item	pass rate	difficulty	discrimination	loading	male d
1	0.60	-0.30	0.61	0.52	0.01
2	0.58	-0.27	0.78	0.62	0.01
3	0.77	-0.84	0.58	0.50	-0.20
4	0.52	-0.05	0.70	0.58	-0.14
5	0.48	0.06	0.66	0.55	0.00
6	0.35	0.46	0.58	0.50	0.10
7	0.48	0.07	0.55	0.48	-0.07
8	0.64	-0.47	0.83	0.64	0.17
9	0.79	-1.02	0.78	0.61	-0.21
10	0.28	0.61	0.30	0.28	0.00
11	0.40	0.29	0.48	0.43	-0.15
12	0.59	-0.26	0.52	0.46	-0.06
13	0.63	-0.43	0.86	0.65	-0.15
14	0.71	-0.76	0.93	0.68	-0.23
15	0.39	0.35	0.79	0.62	-0.01
16	0.40	0.27	0.53	0.47	0.04
17	0.49	0.04	0.82	0.64	-0.11
18	0.57	-0.22	0.88	0.66	-0.07
19	0.48	0.07	0.64	0.54	-0.11

⁷ Pass rates are non-linear functions of the item difficulty and person ability levels (theta). This has the implication that a difference in pass rates is also a non-linear function of the differences in scores between groups and will be affected by the item pass rate. For instance (based on the normal curve), the z score difference between 55% vs. 50% is 0.13 (0.13 vs. 0.00), but the difference between 99% and 94% is 0.77 (2.30 vs. 1.60).

20	0.28	0.64	0.49	0.44	0.02
21	0.55	-0.17	0.80	0.63	0.04
22	0.48	0.04	0.44	0.41	-0.04
23	0.63	-0.44	0.85	0.65	-0.04
24	0.58	-0.26	0.89	0.66	-0.15
25	0.48	0.07	0.75	0.60	-0.15
26	0.33	0.48	0.47	0.42	0.00
27	0.31	0.54	0.44	0.41	0.00
28	0.53	-0.09	0.77	0.61	0.01
29	0.45	0.14	0.63	0.53	0.11
30	0.38	0.38	0.72	0.58	0.00
31	0.52	-0.07	0.74	0.60	-0.09
32	0.67	-0.47	0.43	0.39	-0.24
33	0.61	-0.36	0.80	0.62	-0.16
34	0.68	-0.63	0.91	0.67	-0.13
35	0.67	-0.55	0.75	0.60	-0.10
36	0.63	-0.43	0.76	0.61	-0.18
37	0.73	-0.83	0.92	0.68	-0.26
38	0.67	-0.54	0.74	0.59	-0.26
39	0.38	0.37	0.60	0.52	-0.11
40	0.35	0.46	0.56	0.49	-0.05

Table 1: Item parameters from the main item response theory model. Fit to all data using the 2 parameter normal model.

Both simple scoring and IRT-based scoring were used to score the subjects. The resulting scores correlated at 1.00 (.9996). The IRT-based scores were used for the following analyses.

3.2. Sex differences in scores

Figure 1 shows the fitted distribution of scores by sex.

Figure 1: Distribution of mathematical ability by sex. Density fitting done using the Gaussian kernel (see documentation of `stats::density` for details). Vertical lines show the means.

The test has a slight ceiling problem; 0.50% of the sample obtained perfect scores. There was almost no floor effect, as only 0.05% of subjects obtained raw scores of 0.

Overall, females obtained higher scores, but the difference was very small, $d = 0.12$. Males had very slightly greater variability (male standard deviation 1.02 vs. female 0.98), as has typically been found (Lehre, Lehre, Laake, & Danbolt, 2009).

3.3. Sex differences and item characteristics

Previous studies of tests of mathematical ability have found an item difficulty $\times$ sex performance interaction such that males tend to perform relatively better on the harder items (Beller & Gafni, 2000; Bielinski & Davison, 2001; Penner, 2003). Note that this does not necessarily mean that males have an overall advantage, just that they are relatively more favored/less disfavored on the more difficult items, which might still mean that females do better on them. Figure 2 shows the scatterplot of item difficulty and the male advantage (i.e. negative d means females have an advantage).

Figure 2: Scatterplot of item difficulty and male advantage (negative numbers = female advantage).
CI95 = analytic 95% confidence interval.

A clear relationship can be seen despite girls having the overall advantage ($d = -0.12$). It can also be seen in the figure that item 8 is a strong outlier. Excluding it increases the correlation to 0.68 [CI95: 0.47 0.82]. In this case, an unstandardized measure of effect size is perhaps more useful, so a simple linear model was fit to calculate this. The unstandardized betas were 0.14 and 0.15, respectively for all data and outlier removed subset. For items with difficulties of -1.00, 0.00 and 1.00, the predicted male advantages were -0.22, -0.07, and 0.09 (based on model with the outlier removed). These difficulties correspond to items with pass rates of 81%, 49% and 18% in the current dataset.⁸

The relative lack of difficult items means that the observed female advantage is larger than it would have been if more difficult items had been included. In other words, item-level test composition matters for the size of sex differences. This conclusion has also been recently reached by other researchers examining sex differences in knowledge tests (Schroeders, Wilhelm, & Olaru, 2016). To illustrate the point, the test was split into two halves based on difficulty and each subtest was analyzed as before. The two subtests' IRT scores correlated at .68. The sex differences were -0.15 and -0.07, respectively for the easy and hard subtests.

It has commonly been found that sex differences on cognitive tests depend on the nature of test. In

⁸ These values deviate slightly from the normal distribution predicted values presumably due to the slight non-normality of the scores; kurtosis -0.92, skew = 0.20. The predictions based on the normal curve are 84%, 50% and 16%.

general, some verbal and memory tests favor females, while chronometric, spatial and some mathematical tests favor males (Halpern, 2013; Hunt, 2010). To examine whether such domain-related differences were also found inside the mathematical domain, items were classified on two characteristics: content and Bloom's taxonomy level (Bloom, 1984). The classification of items was done by the test constructors, not the authors.

[examples]

Table 2 show the distribution of item classes.

Type	Count	Percent
<i>Content</i>		
Number	23	57.50
Geometry	12	30.00
Data analysis	5	12.50
<i>Level</i>		
Knowledge	15	37.50
Application	13	32.50
Reasoning	12	30.00

Table 2: Item classification counts.

A regression model was fitted with the outcome being male advantage and the predictors being item position, difficulty, loading, content and level. Results are shown in Table 3.

Predictor	Beta	SE	CI lower	CI upper
Position	0.00	0.00	-0.01	0.00
Difficulty	0.16	0.04	0.08	0.24
Loading	0.20	0.18	-0.16	0.56
Content: number	0.00			
Content: data analysis	-0.06	0.06	-0.18	0.06
Content: geometry	0.01	0.04	-0.06	0.08
Level: knowledge	0.00			

Level: application	0.02	0.03	-0.05	0.09
Level: reasoning	0.02	0.04	-0.06	0.10

Table 3: Model results. Outcome: male advantage. $n=40$ items. $R^2 = .46$, adj. $R^2 = .36$. Reference classes were set to the modes. Unstandardized betas. SE = standard error. CI = 95% analytic confidence interval.

The model confirmed the relationship between difficulty and male advantage with the same effect size as seen in the bivariate analysis. Item content was unrelated to sex differences as was position.

3.4. Measurement bias

Measurement bias is a serious issue for group comparisons (Wicherts, 2016). If the test does not function the same across groups, results of score comparisons may be difficult or impossible to interpret. To examine how similarly the present test functioned for males and females, we re-analyzed the data by sex and compared the item parameters obtained this way. Figures 3-4 show the results for item difficulty and loading.

Figure 3: Item difficulties by sex. Sorted by mean item difficulty.

Figure 4: Item loadings by sex. Sorted by mean item loading.

As can be seen, the item parameters were very similar. Correlations across sex were .97 [.95 to .99] and .90 [.82 to .95] for difficulties and loadings, respectively. Are these larger than one would expect from sampling error alone? To find out, 5 random subsets of half the data (to mirror the subsampling by sex) were created and analyzed. The correlations between the IRT parameters in the random subsets were always at 1.00.⁹ Thus, the observed deviation from 1.00 by sex could not be explained by sampling error and must thus reflect slight deviations from measurement invariance.

To examine whether the small differences in item parameters were related to sex differences, the (absolute) differences between item parameters across sex were correlated with the male advantages. Results are shown in Table 4.

	Male d	Delta difficulty	Delta loading	Abs delta difficulty	Abs delta loading
Male d	1	0.03 [-0.28 0.34]	0.01 [-0.30 0.32]	-0.02 [-0.33 0.30]	-0.05 [-0.36 0.26]
Delta difficulty	0.03 [-0.28 0.34]	1	0.36 [0.05 0.60]	0.74 [0.56 0.92]	0.10 [-0.22 0.42]

⁹ Difficulties are more resistant to sampling error than loadings, but the effect is very small at n=16k. The mean correlation for difficulties was .99969, while for loadings it was .99712.

				0.85]	0.40]
Delta loading	0.01 [-0.30 0.32]	0.36 [0.05 0.60]	1	0.44 [0.15 0.66]	0.12 [-0.20 0.41]
Abs delta difficulty	-0.02 [-0.33 0.30]	0.74 [0.56 0.85]	0.44 [0.15 0.66]	1	0.11 [-0.21 0.41]
Abs delta loading	-0.05 [-0.36 0.26]	0.10 [-0.22 0.40]	0.12 [-0.20 0.41]	0.11 [-0.21 0.41]	1

Table 4: Sex differences in item parameters and performance. $n = 40$. Abs = absolute.

There did not seem to be any relationship between the relative size of the sex difference in item parameters and the sex difference in performance. In other words, the items that functioned the most differently between the sexes did not show directionally larger sex differences. Thus, while the test cannot be said to be completely invariant across sex, the slight invariance seems unrelated to the differences observed. Furthermore, the summary statistics for the loadings were similar by sex, with means of .56 and .55, standard deviations of 0.11 and 0.09, for males and females, respectively.

4. Discussion and conclusion

The present study found multiple results of interest. First, females had a small overall advantage (-0.12 d). This finding is congruent with Richard Lynn's developmental model of sex differences of cognitive ability where females are expected to slightly outperform boys around age 10 (Bakhiet et al., 2017; Colom & Lynn, 2004; Lynn, 2017; Lynn & Kanazawa, 2011). Similar to previous studies (Beller & Gafni, 2000; Bielinski & Davison, 2001; Penner, 2003), an interaction was found between male advantage and item difficulty, seen as a correlation between item difficulty and male advantage of .57 (.68 after removal of 1 outlier). This relationship had the same size in multiple regression with other item characteristics included as controls. This finding is also congruent with Lynn's developmental model because it leads to the prediction that as children age and increase in mathematical ability (and general intelligence), they also encounter items with higher difficulties, which then result in a (larger) male advantage. However, an alternative interpretation is that the easier items load relatively more strongly on reading comprehension, on which females have an advantage. It is not possible to distinguish between these models using the present dataset.

Second, item position has previously been found to be related to relatively better female and lower-scoring group performance. However, it has previously been noted that item position is confounded with item difficulty (Federer, Nehm, & Pearl, 2016). We found no effect of item position in the multiple regression model, in line with the difficulty-confounding interpretation.

Third, measurement invariance with respect to sex was found to be slightly violated for the data. Cross-

sex item correlations for difficulty and loading were .97 and .90, whereas 1.00 was expected based on the sample sizes. However, this violation of measurement invariance did not seem to bias the results in any direction. The item parameter correlations across groups are similar to those seen for e.g. PISA tests (Grisay, Gonzalez, & Monseur, 2009).

Fourth, because the sample used almost an entire population (50% of total), the present results can be said to be extremely reliable with regards to sampling error.

The present study is similarly limited by a number of factors. First, only 40 items were used, which results in somewhat low power to detect small but perhaps important item-level relationships. Related to this, the item distribution among content types and levels was highly imbalanced, reducing the statistical certainty for these analyses.

Second, the study only included mathematical test items, not for instance, verbal, physics or art items. This reduction in item variation in comparison with the universe of all possible items may mean that important relationships might go undetected. Similarly, observed relationships between item parameters in the present study may be due to confounding with the item types present in this study. Future studies should include items of different types to adjudicate between the competing models.

Third, there were very limited information available about the students. For instance, age was not available, and because students can differ slightly in age, this adds noise to the data. If members of one sex were somewhat older than the other sex, this may confound the findings.

Fourth, the sample only contained students from a single grade, which consequently makes it impossible to analyze between grade effects.

Acknowledgments and funding

The authors extend their appreciation to the Deanship of Scientific Research, King Saud University for funding this work through the International Research Group Project RG-1436-028. The funder was not involved with the analyses, writing or submission of this study.

References

- Bakhiet, S. F. A., Abdelrasheed, N. S. G., Cheng, H., Lynn, R., Essa, Y. A. S., & Blahmar, T. A. M. (2017). Sex Differences on the Coloured Progressive Matrices in Sudan. *Mankind Quarterly*, 57(4). Retrieved from <http://mankindquarterly.org/archive/issue/57-4/5>
- Beller, M., & Gafni, N. (2000). Can Item Format (Multiple Choice vs. Open-Ended) Account for Gender Differences in Mathematics Achievement? *Sex Roles*, 42(1–2), 1–21. <https://doi.org/10.1023/A:1007051109754>
- Bielinski, J., & Davison, M. L. (2001). A Sex Difference by Item Difficulty Interaction in Multiple-

- Choice Mathematics Items Administered to National Probability Samples. *Journal of Educational Measurement*, 38(1), 51–77. <https://doi.org/10.1111/j.1745-3984.2001.tb01116.x>
- Bloom, B. S. (1984). *Taxonomy of Educational Objectives Book 1: Cognitive Domain* (2nd edition edition). New York: Addison Wesley Publishing Company.
- Colom, R., & Lynn, R. (2004). Testing the developmental theory of sex differences in intelligence on 12–18 year olds. *Personality and Individual Differences*, 36(1), 75–82. [https://doi.org/10.1016/S0191-8869\(03\)00053-9](https://doi.org/10.1016/S0191-8869(03)00053-9)
- DeMars, C. (2010). *Item response theory*. Oxford ; New York: Oxford University Press.
- Federer, M. R., Nehm, R. H., & Pearl, D. K. (2016). Examining Gender Differences in Written Assessment Tasks in Biology: A Case Study of Evolutionary Explanations. *CBE-Life Sciences Education*, 15(1), ar2. <https://doi.org/10.1187/cbe.14-01-0018>
- Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI*, 2, 63–83.
- Halpern, D. F. (2013). *Sex differences in cognitive abilities*. Psychology press. Retrieved from <https://books.google.com/books?hl=en&lr=&id=ocl5AgAAQBAJ&oi=fnd&pg=PP1&dq=halpern+sex+differences&ots=WCaMyeundr&sig=IKuHrR8os0BE43Yfj8Y3AE5Lt3A>
- Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The Science of Sex Differences in Science and Mathematics. *Psychological Science in the Public Interest : A Journal of the American Psychological Society*, 8(1), 1–51. <https://doi.org/10.1111/j.1529-1006.2007.00032.x>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, 33(2–3), 61-83-135. <https://doi.org/10.1017/S0140525X0999152X>
- Hunt, E. (2010). *Human Intelligence*. Cambridge University Press.
- Irwing, P., & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. *British Journal of Psychology*, 96(4), 505–524. <https://doi.org/10.1348/000712605X53542>

- Lehre, A.-C., Lehre, K. P., Laake, P., & Danbolt, N. C. (2009). Greater intrasex phenotype variability in males than in females is a fundamental aspect of the gender differences in humans. *Developmental Psychobiology*, 51(2), 198–206. <https://doi.org/10.1002/dev.20358>
- Lynn, R. (2017). Sex Differences in Intelligence: The Developmental Theory. *Mankind Quarterly*, 58(1). Retrieved from <http://mankindquarterly.org/archive/issue/58-1/2>
- Lynn, R., & Kanazawa, S. (2011). A longitudinal study of sex differences in intelligence at ages 7, 11 and 16 years. *Personality and Individual Differences*, 51(3), 321–324. <https://doi.org/10.1016/j.paid.2011.02.028>
- Penner, A. M. (2003). International gender $\times$ item difficulty interactions in mathematics and science achievement tests. *Journal of Educational Psychology*, 95(3), 650–655. <https://doi.org/10.1037/0022-0663.95.3.650>
- R Core Team. (2017). R: A Language and Environment for Statistical Computing (Version 3.4.0). R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org>
- Revelle, W. (2017). psych: Procedures for Psychological, Psychometric, and Personality Research (Version 1.7.8). Retrieved from <https://cran.r-project.org/web/packages/psych/index.html>
- Schroeders, U., Wilhelm, O., & Olaru, G. (2016). The influence of item sampling on sex differences in knowledge tests. *Intelligence*, 58, 22–32. <https://doi.org/10.1016/j.intell.2016.06.003>
- Uebersax, J. S. (2015, September 8). Introduction to the Tetrachoric and Polychoric Correlation Coefficients. Retrieved from <http://john-uebersax.com/stat/tetra.htm>
- Whitley, E., Deary, I. J., Ritchie, S. J., Batty, G. D., Kumari, M., & Benzeval, M. (2016). Variations in cognitive abilities across the life course: Cross-sectional evidence from Understanding Society: The UK Household Longitudinal Study. *Intelligence*, 59, 39–50. <https://doi.org/10.1016/j.intell.2016.07.001>
- Wicherts, J. M. (2016). The importance of measurement invariance in neurocognitive ability testing. *The Clinical Neuropsychologist*, 30(7), 1006–1016. <https://doi.org/10.1080/13854046.2016.1205136>
- Wicherts, J. M. (n.d.). Psychometric problems with the method of correlated vectors applied to item

scores (including some nonsensical results). *Intelligence*.

<https://doi.org/10.1016/j.intell.2016.11.002>

Wicherts, J. M., & Johnson, W. (2009). Group differences in the heritability of items and test scores.

Proceedings of the Royal Society B: Biological Sciences, 276(1667), 2675–2683.

<https://doi.org/10.1098/rspb.2009.0238>