Sex differences in arithmetical performance scores: Central tendency and variability

R. Martens a,*, P.P.M. Hurks a, C. Meijs b, R. Wassenberg c, J. Jolles d

a Faculty of Psychology and Neuroscience, Maastricht University, P.O. Box 616, NL – 6200 MD, Maastricht, The Netherlands
b Ruud de Moor Center, Open University, P.O. Box 2960, NL – 6401 DL, Heerlen, The Netherlands
c Department of Psychiatry and Psychology, Maastricht University Medical Center, P.O. Box 5800, NL – 6202 AZ, Maastricht, The Netherlands
d Faculty of Psychology & Education, VU University Amsterdam, Van der Boechorststraat 1, NL – 1081 BT, Amsterdam, The Netherlands

A R T I C L E   I N F O

Article history:
Received 15 October 2010
Received in revised form 13 April 2011
Accepted 10 June 2011

Keywords:
Arithmetical abilities
Sex differences in development
Central tendency
Variability
Precociousness

A B S T R A C T

The present study aimed to analyze sex differences in arithmetical performance in a large-scale sample of 390 children (193 boys) frequenting grades 1–9. Past research in this field has focused primarily on average performance, implicitly assuming homogeneity of variance, for which support is scarce. This article examined sex differences in arithmetical operations (addition, subtraction, multiplication and division) in central tendency and variability. Central tendency analyses revealed a male advantage, predominantly in addition and subtraction, from grade 6 onward. Variability analyses showed that sex differences were largest among higher achievers and absent among lower achievers. Thus, central tendency and variability analyses provide complementary information on sex differences in arithmetical performance. In conclusion, sex differences in arithmetical performance exist, but depend on the studied arithmetical operation, age group and achievement level. The present study thereby offers new directions for future research by indicating the need for a broader perspective on sex differences.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

In many industrialized countries in the world, women are increasingly dominating specific disciplines, such as law, medicine and psychology. In contrast, women continue to be underrepresented in mathematics, science, and engineering. Even in high school and college, this pattern of sex related interests is assumed to be present (Halpern, 2004; Halpern et al., 2007). Why are mathematics and science still dominated by males, while other disciplines have become progressively female-dominated? At the heart of this question lies the discussion if and how males and females differ innately in terms of, e.g., their mathematical skills. Studying sex differences across age may provide information about the robustness of such differences.

Questions about sex differences in the cognitive and scholastic abilities needed for success in mathematics have a long history in scientific psychology (Halpern et al., 2007). For instance, some researchers have studied sex differences in terms of arithmetical abilities (such as addition and multiplication). The levels of arithmetical abilities that children attain early in school seem to be predictive for more advanced mathematics later in life (e.g., Geary, 1994; Geary, Frensch, & Wiley, 1993; Gersten, Jordan, & Flojo, 2005). However, studies focusing on sex differences in arithmetic are limited and inconsistent in their outcome. Where some studies have failed to find sex differences in overall arithmetical abilities (e.g., Hyde, Fennema, & Lamon, 1990; Imbo & Vandierendonck, 2007; Lachance & Mazzocco, 2006), others have reported sex differences in mean arithmetical test scores, primarily favoring males (e.g., Geary, Saults, Liu, & Hoard, 2000; Lynn & Irwing, 2008; Royer, Tronsky, Chan, Jackson, & Marchant, 1999).

Most of these studies have focused primarily on male–female differences in average performance (i.e., central tendency) on arithmetic tests. This type of research implicitly assumes homogeneity of variance, even though support for the validity of that assumption is scarce. Studies reporting performance variability provide complementary information. Hedges and Nowell (1995) found, for instance, that in addition to (small) mean differences in e.g., mathematical and science abilities favoring older adolescent males, male performance was more variable than female performance, with variance ratios primarily between 5 and 28%. Strand, Deary, and Smith (2006) found that male 11- and 12-year-olds dominated the top stanines of both quantitative and nonverbal reasoning tests. For instance, in the 9th stanine (i.e., top 4% of all scores) of a quantitative reasoning test 60% was male. Thus, primarily at the top end of the distribution substantial sex differences in mathematical abilities are
found (Halpern et al., 2007; Hyde et al., 1990). Unfortunately, these previous studies focused primarily on adolescents. To the best of our knowledge, only one study considered the variability of arithmetical scores in childhood. More specifically, Royer et al. (1999) reported on sex differences in the distribution of arithmetical abilities across grades, while focusing in particular on the high-end of the distribution. They found that in grades 5–8 the best achieving males were faster and more accurate than the best achieving females, while there was no clear difference between the best achieving males or females in grades 1–4 or the lowest achieving males or females in grades 1–4 and 5–8. These findings suggest that sex differences in arithmetical performance depend on both achievement level and age, making this method for studying sex differences very promising. However, although the mean scores for both males and females in the slow and fast groups were mentioned separately in this study, the level of significance and the effect sizes of the reported sex differences were not. This makes it difficult to assess how large the effects of these differences were. Also, Royer et al. (1999) only included a selection of arithmetical operations in their study (i.e., not including division). Given that previous studies have reported differential representations of arithmetical operations in the brain (e.g., Dehaene & Cohen, 1995; Dehaene, Piazza, Pinel, & Cohen, 2003), it may be that sex differences vary as a function of type of arithmetic operation (e.g., addition vs. multiplication).

Hence, we aimed at increasing the knowledge on sex differences in arithmetical performance by reporting both significance levels and effect sizes and combining all factors discussed above: i.e., including all arithmetical operations separately (addition, subtraction, multiplication and division), an age range of 6–15 years, and multiple achievement level measures (central tendency and variability, with a particular interest in sex differences at the high end of the distribution). Apart from these factors, item difficulty (Bielinski & Davison, 2001; Penner, 2003) and test format (i.e., multiple choice vs. open-ended; DeMars, 2000; Wester & Henriksson, 2000) could also be considered as possible influences on the direction and magnitude of sex differences in mathematical abilities. To minimize the possible confounding influences of these factors in the present study, all participants completed the same time-restricted arithmetical abilities test (i.e., the Arithmetic Tempo Test; De Vos, 1992).

2. Method

2.1. Procedure and participants

This study uses data of a large-scale, cross-sectional and longitudinal research program that focuses on mechanisms underlying cognitive development (e.g., Hurks et al., 2010; Wassenberg et al., 2008). In the present study, participants came from multiple schools for regular elementary (Nschools = 26) and secondary education (ranging from lower secondary professional education to pre-university education, Nschools = 5) in the Netherlands. Participating schools were asked to distribute information packages among caregivers of all children in the first (i.e., 6–7 years), third, fifth, and sixth grades of elementary school, and seventh, eighth, and ninth grades of secondary school. Besides information about the purpose of the study, the information packages contained a stamped return envelope, an informed consent letter, and a questionnaire on characteristics of the child and his environment, which caregivers completed in if they agreed to participate. After consent was obtained, all children were screened on the bases of exclusion criteria (see Hurks et al., 2010; i.e., not being in the appropriate grade, not having the Dutch nationality, and current use of medication that could affect cognitive functioning. Developmental disorders, such as learning disorders, were not considered to be an exclusion criterion, thereby preventing the sample from being “overly normal”. Instead, children were considered to develop at a normal pace when they attended a school for regular education and were in the appropriate grade for their age. After screening, only children meeting the predetermined requirements were approached for further testing.

In total, 390 children aged 6 to 15 (193 boys) participated in the present study. In Table 1, the descriptive statistics of all children are given per grade. All selected children completed a fixed-order neuropsychological test battery concentrating on various cognitive domains, such as arithmetical abilities, working memory and estimation. Four well-trained research assistants tested all children. Testing took place in stimulus-free rooms at the participating schools and took approximately 90 min. The Ethics Committee of the Faculty of Psychology and Neuroscience of Maastricht University approved the research protocol.

2.2. Instruments

2.2.1. Arithmetical abilities

The Arithmetic Tempo Test (ATT; De Vos, 1992) is a standardized paper-and-pencil test frequently used in Dutch and Flemish education to measure arithmetical abilities. Its psychometric value has been established using a sample of 10,059 children (Desoete, Stock, Schepens, Baeyens, & Roeyers, 2009; Stock, Desoete, & Roeyers, 2009). Four sets of 40 arithmetical problems were presented to the children, respectively containing addition (ATT 1, +), subtraction (ATT 2, −), multiplication (ATT 3, ×), and division problems (ATT 4, ÷). Per set, the children have to solve as many problems as possible within 1 min. All problems consisted of two-operand equations with an outcome smaller than 100 and both

<table>
<thead>
<tr>
<th>Variables</th>
<th>1 (N = 65)</th>
<th>3 (N = 60)</th>
<th>5 (N = 60)</th>
<th>6 (N = 77)</th>
<th>7 (N = 45)</th>
<th>8 (N = 49)</th>
<th>9 (N = 34)</th>
<th>Statistic F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.41*</td>
</tr>
<tr>
<td>Boys</td>
<td>28%</td>
<td>30%</td>
<td>27%</td>
<td>34%</td>
<td>21%</td>
<td>28%</td>
<td>21%</td>
<td></td>
</tr>
<tr>
<td>Girls</td>
<td>37%</td>
<td>30%</td>
<td>38%</td>
<td>43%</td>
<td>53%</td>
<td>43%</td>
<td>38%</td>
<td></td>
</tr>
<tr>
<td>M (SD)</td>
<td>30 (14)</td>
<td>29 (13)</td>
<td>31 (15)</td>
<td>33 (16)</td>
<td>37 (19)</td>
<td>37 (19)</td>
<td>34 (17)</td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td>10.94 (2.86)</td>
<td>10.44 (3.16)</td>
<td>10.51 (2.61)</td>
<td>10.15 (3.16)</td>
<td>8.93 (2.10)</td>
<td>10.20 (2.03)</td>
<td>9.53 (2.14)</td>
<td>1.65**</td>
</tr>
<tr>
<td>Age at measurement</td>
<td>7.36 (0.32)</td>
<td>9.36 (0.34)</td>
<td>11.43 (0.35)</td>
<td>11.68 (0.58)</td>
<td>13.50 (0.31)</td>
<td>14.38 (0.33)</td>
<td>15.52 (0.33)</td>
<td>2710.09***</td>
</tr>
</tbody>
</table>

1 = 1st grade; 3 = 3rd grade; 5 = 5th grade; 6 = 6th grade; 7 = 7th grade; 8 = 8th grade; 9 = 9th grade; vocabulary = estimate of verbal intelligence, measured with the vocabulary subtest of the Dutch Wechsler Intelligence Scales Revised; dash indicates that data was not available.

* One-way analysis of variance (df1 = 6, df2 = 383).
** p < .001.
operands ranging between 0 and 90. Test scores were calculated per set, i.e., the total amount of problems answered correctly minus the total amount of problems answered incorrectly. Additionally, an overall score (ATT Total) was calculated consisting of the sum of ATT 1–4.

2.2.2. Verbal intelligence

To have an estimate of verbal intelligence (VIQ), the WISC-R Vocabulary subtest (De Bruyn et al., 1986; Wechsler, 1974) was administered. Children were asked to give a definition of words (ranging from easy to difficult) as accurately as possible. Scaled scores ranged from 1 to 19 ($M = 10, SD = 3$). Because the present study had already started when the WISC-III became available in the Netherlands, the WISC-R was used (Wassenberg et al., 2008). The reliability coefficient for the Vocabulary subtest in a Dutch population of children is .86–.88 (De Bruyn et al., 1986).

2.3. Statistical analyses

All analyses were performed using SPSS 16.0 for Macintosh OS X. Performance on all ATT outcome measures was equal among the four test administrators (F-values ranging from 0.08 to 1.65). For seven children the ATT data (or part of it) were missing or unreliable, due to refusal to cooperate (e.g., because of fatigue) and/or failure to execute a task according to instructions. Missing data were less than 5% of the total data; therefore, these were not replaced (Croy & Novins, 2005). Unreliable data were excluded from later analyses. No extreme values (i.e., more than three times the interquartile distance <25th percentile or >75th percentile as defined by Huizingh, 2002) were found on the ATT outcome measures for any of the other children.

Two additional comments should be made. Firstly, in line with ATT guidelines (De Vos, 1992), all children in first grade only completed ATT 1 and 2. This was a deliberate choice, since multiplication (ATT 3) and division (ATT 4) are not yet taught at this age in the Dutch school system. As a consequence, scores of children from grade 1 could not be included in any of the Multivariate analyses containing ATT 3, ATT 4 or ATT Total scores (see below). Secondly, data on verbal intelligence were lacking for all children in grade 6. However, univariate analyses of variance (ANOVA) revealed no significant differences in VIQ scores between all other grades (see Table 1).

Using GLM Multivariate analyses, central tendencies measures of ATT 1–4 as well as ATT Total were analyzed as a function of grade and sex. Next, the participants were divided into two groups based on the grades they frequented: i.e., (a) grades 1, 3 and 5 vs. (b) grades 6 to 9.

### 3. Results

#### 3.1. Arithmetical abilities, in terms of central tendency, across grades

Fig. 1 shows the mean ATT performance for addition (ATT 1), subtraction (ATT 2), multiplication (ATT 3), and division (ATT 4) across grades. GLM Multivariate analyses (without grade 1) including all outcome measures revealed a significant main effect for grade, Pillai’s Trace: $F(20, 1232) = 13.01, p < .001$. The same holds true for GLM Multivariate analyses including only ATT 1 and 2 (i.e., with grade 1), Pillai’s Trace: $F (12, 742) = 38.90, p < .001$. Results showed that, independent of type of operation tested, arithmetical performance improved significantly until grade 5, with each grade performing better on average than the previous one ($p < .001$ for comparisons of grades 1–5). Thereafter, a slight increase in performance could still be observed until grade 8, however, this increase was less strong and less consistent (i.e., ATT 1: grade 6 vs. grade 7: $p < .01$, grade 7 vs. grade 8: $p < .05$, ATT 2: grade 7 vs. grade 8: $p < .01$; ATT 3: grade 7 vs. grade 8: $p < .05$, ATT 4: grade 7 vs. grade 8: $p < .05$; ATT Total: grade 7 vs. grade 8: $p < .01$).

In addition, a significant differential effect was found for the type of arithmetical operation: addition > subtraction > multiplication > division (# correct — incorrect; for all univariate analyses: $p < .001$). This significant difference in development trajectories between the four arithmetical operations was confirmed by a GLM Repeated Measures analysis; Pillai’s Trace: $F (15, 942) = 7.45, p < .001$.

#### 3.2. Sex differences (and sex × age interactions) in arithmetical abilities in terms of central tendency

A GLM Multivariate analysis of all ATT performances indicated a trend towards a main effect of sex: i.e., boys tended to outperform girls in arithmetical performance independent of grade, $F (4, 305) = 2.48, p = .04$ (see Fig. 2). Post-hoc Bonferroni comparisons revealed that this finding could primarily be explained by sex differences in

![Fig. 1. Arithmetical performance across grades shown per arithmetical operation.](image-url)
ATT 1 (addition) and ATT 2 (subtraction) scores, respectively. The overall scores for ATT 1 were: Mboys = 25.84 (SD = 8.12) vs. Mgirls = 24.24 (7.68), and for ATT 2: Mboys = 22.92 (8.11) vs. Mgirls = 20.62 (8.06). Separate cross-sectional comparisons of ATT 1 and 2 for all grades (including grade 1), using GLM univariate analyses, revealed similar results.

The GLM Multivariate analysis also revealed a trend towards an interaction between grade and sex, Pillai’s Trace: F (20, 1232) = 1.53, p = .06. Post-hoc Bonferroni comparisons showed that this trend was primarily caused by a grade × sex interaction for ATT 1 performance (i.e., a small male advantage was found only from grade 6 onward), Pillai’s Trace: F (5, 308) = 2.31, p = .04. Two additional separate GLM Multivariate analyses of ATT performance for either grades 3–5 or grades 6–9, with sex as independent variable revealed similar results. In the analysis of grades 3–5 no significant sex differences were found, while the analysis of grades 6–9 indicated a trend towards a male advantage, Pillai’s Trace: F (4, 196) = 2.69, p = .03. Post-hoc Bonferroni comparisons revealed that, in grades 6–9, sex differences were significant only for ATT 1 and 2; resp. F (1, 199) = 7.95, p < .01, d = 0.38 and F (1, 199) = 8.58, p < .01, d = 0.42. A trend was found for ATT Total, F (1, 199) = 5.77, p = .02, d = 0.34.

In sum, sex differences were found primarily in older children performing addition and to a lesser extent subtraction. In contrast, no effects of sex or grade × sex interactions were found on tasks measuring the ability to conduct multiplication and division.

3.3. Sex differences in low, intermediate, and high arithmetical performance: measuring variability

In order to study sex differences in terms of the variability of ATT scores, the performance of low-, intermediate-, and high-scoring participants was analyzed independent of grade, using GLM Univariate analyses. Sex differences were found on all outcome measures within the high-scoring group favoring boys (F-values varied between 16.44 and 37.36, p-values < .001, d-values ranged from 0.31 to 0.54; see Fig. 3 for a reproduction of the pattern of sex differences on the ATT Total, which reflects the pattern seen on all ATT measures). For the intermediate-scoring group, some sex differences were also found, again favoring boys. However, these differences were only significant for ATT 1 and 2 scores (i.e., addition and subtraction), resp. F (1, 191) = 17.69, p < .001, d = 0.26 and F (1, 190) = 22.02, p < .001, d = 0.33 and ATT Total F (1, 155) = 7.60, p < .01, d = 0.23. For the low-scoring participants, no sex differences were found on any ATT measure.

To study grade × sex interactions in this context, two groups were formed, i.e., children frequenting grades 1–5 vs. children frequenting grades 6–9. When studying performance in grades 1–5, no sex differences were found in the low-, intermediate-, or high-scoring groups. Conversely, when analyzing the performance in grades 6–9, sex differences were found in the intermediate-, and high-scoring groups, again favoring boys. These results are similar to the results of the overall analyses of all grades: in the intermediate group significant differences were found for ATT 1 and 2 scores, resp. F (1, 106) = 11.78, p = .001, d = 0.63 and F (1, 106) = 12.57, p = .001, d = 0.65, as well as a trend for ATT Total, F (1, 102) = 5.05, p = .03, d = 0.43. In the high performance group, the differences were significant for all ATT-measures (F-values varied between 16.44 and 24.30, p-values < .001, d-values ranged from 1.12 to 1.23).

In sum, taking variability scores into account, sex differences were primarily found in older, high- and to some extent intermediate-scoring children.

4. Discussion

The influence of sex differences on arithmetical performance in children aged 6 to 15 years was studied in relation to arithmetical operation (i.e., addition, subtraction, multiplication and division), grade, and achievement level (i.e., central tendency and variability). With respect to the first factor, i.e., type of arithmetical operation, it should be noted that most studies reporting on sex differences in arithmetical performance do so on the basis of general achievement scores (e.g., Imbo & Vandierendonck, 2007; Lynn & Irwing, 2008; Rosselli, Ardila, Matute, & Inozemtseva, 2009). Although the results of the limited number of studies into age-related sex differences in arithmetic were promising (i.e., Rosselli et al., 2009; Royer et al., 1999), these studies did not examine all four basic arithmetical operations separately. In the present study, the influence of sex differences on arithmetical abilities was examined more in detail by looking at sex differences per arithmetical operation (addition, subtraction, multiplication and division). On a behavioral level, children improved on all four arithmetical operations up to at least grade 8 (age 13–14 years), with the largest improvement up to grade 5 (age 10–11 years). More importantly, however, the results revealed significant differences in the development trajectories of the arithmetical operations (i.e., total number correct minus incorrect: addition > subtraction > multiplication > division). In line with this, previous studies have reported differential representations of arithmetical operations in the brain (e.g., Dehaene & Cohen, 1995; Dehaene et al., 2003) as well as differences in brain activation patterns when learning these operations (Ischebeck et al., 2006). These studies support the necessity of analyses test performance separately per operation, when studying sex differences in arithmetical performance, as is done in the present study.

Fig. 2. Sex differences in arithmetical performance (ATT scores) shown per arithmetical operation.

Fig. 3. Sex differences in ATT Total scores shown per achievement level (low-intermediate-high).
More importantly, the present study showed that there are sex differences in arithmetical performance, yet that these varied not only with the arithmetical operation being studied, but also with age and achievement level (i.e., variability vs. central tendency). More specifically, while using central tendency scores, sex differences (favoring boys) were found primarily in older children performing addition and to a lesser extent subtraction. In contrast, no effects of sex or grade × sex interactions were found on tasks measuring the ability to conduct multiplication and division. The inclusion of variability scores, however, led to a subtler picture: sex differences (again favoring boys) were primarily found in older, high- and to some extent intermediate-scoring children. This emphasizes the relevance of simultaneously exploring the interaction between the type of operation, grade and achievement level to unravel the complex issue of sex differences.

These findings are in line with the only study that has previously included similar factors while relating sex differences to arithmetical abilities (i.e., Royer et al., 1999). However, the sample sizes in the study by Royer et al. were too small for statistical analyses; their results were based on observational comparisons of mean scores. Consequently, making inferences solely on the basis of Royer et al.’s (1999) results is difficult. The present study reported both the significance levels and the effect sizes of sex differences in arithmetical abilities per performance group, thereby providing more information about the extent of such differences. Our effect size calculations revealed that sex differences in terms of central tendency were small. In contrast, sex differences between high performers were accompanied by predominantly medium effect sizes for all grades and even very large effect sizes when only taking grades 6 to 9 into account. These findings further support the relevance of including variance scores when studying sex differences in children; i.e., small main effects can indeed conceal larger and more significant differences at the top end of the performance distribution.

The found results may well be explained theoretically in terms of inter-individual differences in childhood use of fact retrieval vs. other strategies (such as rehearsal) to solve arithmetical tasks. For one, addition and subtraction are learned first in elementary school, while at least 1 year later formal schooling of multiplication and division starts. As a consequence, addition and subtraction may be better automatized in older children, causing fact retrieval to be used more frequently as a solution strategy. Conversely, if multiplication and division are learned later and are less automatized, even older children may prefer to use other solution strategies (i.e., overt strategies). Given the fact that in the present study sex differences were found primarily on addition and subtraction tasks, one might conclude that boys outperform girls in terms of more automatized fact retrieval, as was proposed by e.g., Royer et al. (1999). The time-restricted arithmetical abilities test that was used in the present study may have stimulated the use of fact retrieval more than a non-time-restricted test, since performance is based on a trade-off between speed and accuracy.

Although the findings of the present study are interesting, there are also some limitations. Firstly, the present study only included one type of arithmetical abilities measure (i.e., arithmetical fact retrieval task) and a quantitative method for scoring. Using additional arithmetical tasks, such as arithmetical word problems and a complementary qualitative scoring method, such as a measure of strategy-use (as suggested by Carr, Jessup, & Fuller, 1999), may provide an even more complete insight into sex differences in arithmetical performance. Secondly, although the present study’s method for variability analyses is promising and may shed new light on sex differences in both arithmetical and other abilities, there is still no consensus on the correct manner of performance selection and which measures to use for this selection (i.e., reaction time or accuracy scores). Consequently, the method of analysis and type of measure used in the present study for dividing the participants into low, intermediate and high performers is different from the one used by for instance Royer et al. (1999). Finally, although our findings are consistent and promising, some prudence is in order due to the relatively low numbers of participants per group. Future research should expand this line of research using an even larger number of participants. This in turn may facilitate the application of more advanced statistical methods, such as multi-level analyses, with the additional benefit of ruling out possible effects of educational differences on our findings. Given the relatively small sample sizes per group, the influence of education-related factors could not be completely excluded. Further, it would be interesting to expand this line of research to older adolescents and adults, to explore if this leads to similar results or if sex differences in multiplication and division arise in older age groups.

5. Conclusion

The present study revealed differences in variance between boys and girls on arithmetical performance, with the most significant sex differences favoring high performing boys in grades 6 to 9. To the best of our knowledge, we are one of the first to study this combination of factors in a developmental context, thereby contributing significantly to the ongoing discussion of sex differences in terms of cognitive performance and educational achievement.

References


