A statistical enquiry into the effect of test length on a math gender gap

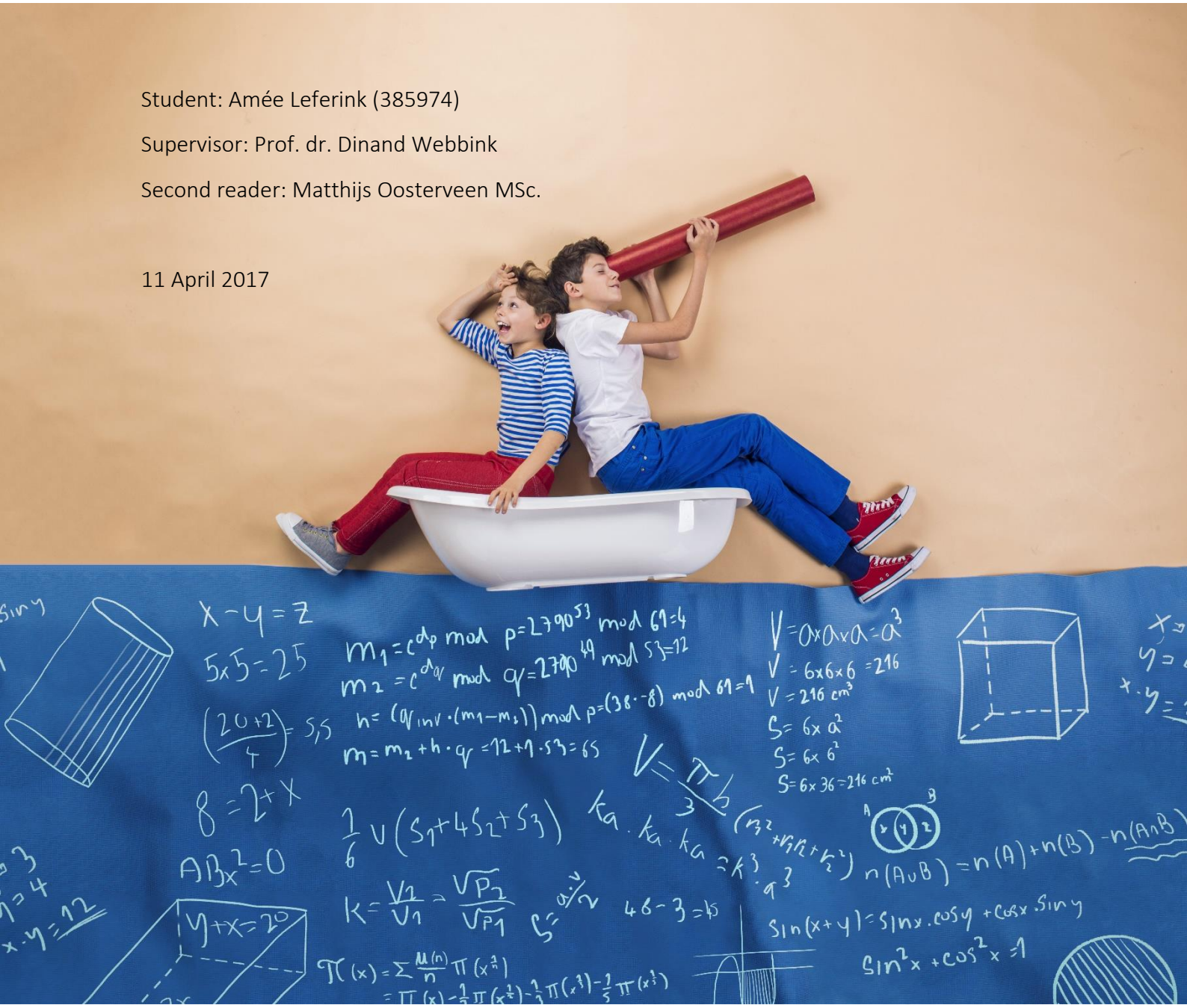
Thesis submitted in partial fulfilment of the requirements for the degree of MSc. Policy Economics from the Erasmus University Rotterdam

Student: Amée Leferink (385974)

Supervisor: Prof. dr. Dinand Webbink

Second reader: Matthijs Oosterveen MSc.

11 April 2017



Foreword

Writing this thesis was both an educative and a lengthy process. Being aware from the start that due to other activities I would not be able to graduate before April the academic year after I commenced, a challenge throughout was prioritizing working on this thesis. I believe this wider timeframe brought with it the advantage of being able to work in an iterative way. I gained insights in working with STATA and I learned to do a large assignment by myself. In searching for an improvement of my style of work, tools I did not use before such as Trello, LucidChart and Todoist were of great help to me. Retaking courses and adding extra ones helped me to make progress and I found academic, moral and technical support from my direct and indirect surroundings.

First and foremost, I received many invaluable insights from my supervisor Prof. Dr. Dinand Webbink and second reader Matthijs Oosterveen MSc., the latter of whom went beyond his formal requirements in giving advice throughout the process. Both have guided me with particularly good humour and helpful instructions, while allowing me to try things out on my own. For their patience and kind attitude I am very grateful. I am moreover appreciative of Prof. Dr. Hyde who agreed to share her datafile, easing the process tremendously as I could build upon her work. Also the contact with researchers, teachers and civil servants, who were kind enough to help me filling the final data set meant a positive surprise regarding the amount of kindness in academia: something I will not soon forget.

I am moreover indebted to the technical and moral support of friends wide and far. My twin sister Tessa proof-read from start to end and cheered me on whenever slightly required, my love Lars was there for me with his ever generous motivation and kindness. I am very much indebted to my friends Marloes with her perfect English and Wim with his econometric background for their technical advice. My parents, Marina and Ben, whom have supported me not just throughout this thesis but throughout the last odd 25 years – I am unquestionably very thankful for both their advice and practical support.

I accept the full responsibility for all ideas in this thesis and any errors or mistakes that it may contain. I have tried to reference and give credit properly where due. Finally, I would like to dedicate this thesis not to a person, but primarily to the two ideas my years at university have taught me: that no one holds a monopoly on truth and that everything can be learned.

Abstract

This study researches whether girls are advantaged (vis-à-vis boys) as math test length increases. The meta-analysis on trends in gender and mathematics performance done by Lindberg, Hyde, Petersen & Linn (2010), was extended with variables on test length to obtain the final dataset for analysis. This contains 435 studies, in which 1,277,598 subjects are included. Proxies for test length are the number of questions and maximum time allowed in the studies. Small negative coefficients for both length measures are found when regressing test lengths on the standardized score difference, in the range of -.0007 to -.0022. The full regression model including test and student background controls and weighted observations estimates a negative coefficient of -.002 ($p<0.05$) for the number of questions and -.002 ($p=0.15$) for the maximum amount of minutes on a standardized difference. These findings indicate performance bias may be a threat for tests as an evaluation mechanism. However, further sensitivity analyses yields mixed results. Combined with the explorative character of this thesis the need for further research to test the hypothesis of gender bias with increasing test length is evident.

Table of Contents

1 Introduction

Most educational systems in the world use standardized tests to keep track of progress and stimulate performance of students and schools alike. The subject of testing has always been precarious as by the design of the tests, some groups may be advantaged or disadvantaged vis-à-vis others. What if certain groups of examinees are advantaged or disadvantaged by a test's design? Does the examination still provide a realistic indication of someone's knowledge or skills? As the length of a test may threat the validity of the evaluation mechanism, this thesis focuses on mathematics tests and whether as test length increases, one gender is preferred over the other.

Over the last decades, psychologists, economists and educationalists have researched the performance of children in different school subjects. Tests in mathematics have been regarded of particular importance because of their indication of future educational success and earnings (Rose & Betts, 2004, a.o). Moreover, there are indications that a larger share of the population educated in the STEM fields[1] corresponds with increased national and technological innovation (Hanushek & Woessmann, 2008, a.o.). As investing in mathematic skill development can be valuable to society, an effective assessment of students' level of math is needed to keep track.

Of particular interest is the debated male pre-eminence in math – causing the so-called math gender gap. Since individual economic advantages seem to be associated with math skills, particular interest for the genders' different responses to different test designs is justified. The design of tests is important, as students may fail to demonstrate their abilities if the exam format does not adequately test their knowledge of a subject. Different designs could in themselves be more compelling to different subgroups. Previous research on math test design has focused on the type of answering, i.e. open versus multiple choice options, and specific aspects of mathematics. Test length and timing, however, has only been limitedly researched.

The discussion of this issue was instigated when Balart & Oosterveen (2017) looked into test results of boys and girls. They assessed 2009 PISA (Programme for International Student Assessment) tests and concluded that with questions posed towards the end of the math section, the difference in favour of boys was significantly smaller than with the questions at the start. The PISA test is a low-stakes test that contains 50-70 questions with 90 minutes assigned for mathematics. Balart & Oosterveen (2017) found that the within-test performance decline for boys was more severe than that of girls. This could be an indication that as tests last longer, boys score relatively low compared to girls – which would mean a downward bias of the actual math gender gap when tests are lengthier. These results may, however, only exist within tests: in any test (regardless of its length), towards the end boys' results are on average lower than at the start, whereas girls manage to keep up a similar performance across the full length. The latter would indicate that in any (math) test, boys' actual skill level may be higher than test results suggest. This research sparked the interest for the broader hypothesis examined in this thesis: *Is there female favoured test bias as test length increases?*

This thesis is to be considered an explorative research on the topic. The dataset of an original meta-analysis is expanded with test length variables. With data of these studies; their test lengths; and found gender difference in math tests, I will then assess the correlation between test length and performance. Three aspects to answer the hypothesis are considered of relevance:

[1] Science, technology, engineering and mathematics - all fields are related to mathematics

1. What is the influence of a test's number of questions on gender differences in math scores?
2. What is the influence of a test's maximum test-time allowed on gender differences in math scores?
3. Are there specific student or test characteristics that impact a test length effect and make it more or less profound?

In order to answer these questions, the widely-cited meta-analysis on children's and adults' math tests published in the Psychological Bulletin in 2010 by Lindberg, Hyde, Petersen and Linn is used. Their dataset includes observations from 242 peer-reviewed articles, equalling to 441 studies. It was expanded to include test length in minutes and the number of questions per study.

This thesis finds mixed results of the effect of test length on the math gender gap. Small negative coefficients for both length measures are found when regressing test lengths on the standardized score difference, in the range of -.0007 to -.0022 per additional minute or question. The full regression model including test and student background controls and weighted observations estimates a negative coefficient of -.002 ($p<0.05$) for the number of questions and -.002 ($p=0.15$) for the maximum amount of minutes on a standardized difference. However, further sensitivity analyses yield mixed results.

Reader's guide

This thesis is structured as follows: in the coming section (2) of this thesis, I look into the broader academic literature around testing and the math gender gap. The first section is a literature review on the found gender gaps in math tests (2.1), different explanations for both a gap and the mixed findings (2.2) and economic relevance of the math gender gap (2.3.). The potential importance of test design is considered thereafter (2.4). An elaboration on the data collection and a data description is provided in section (3). This section is split up in four subsections: (3.1) data collection, (3.2) summary findings by Lindberg, Hyde, Petersen & Linn, (3.3) extension of the dataset and (3.4) description of statistics. Section (4) presents the empirical strategy and method and considers internal validity (4.1) and missing observations (4.2 and 4.3). I then consider the equality of test and student characteristics of short and long tests (4.4). The data analysis is done in section (5), which consists of (5.1) regression estimates and (5.2) further sensitivity analysis. The final section (6) includes the conclusions (6.1), limitations and suggestions for further research (6.2) and policy implications of the findings (6.3). At the end the bibliographies (7) and appendices (8) can be found.

2 Literature review

A gender gap in mathematics has been investigated in varying age groups across academic fields over the last few decades. Girls perform better on tests in most school subjects, such as languages, history and reading (see e.g. OECD, 2014; Pekkarinen, 2012). This gender gap, however, is reversed for mathematics and the natural sciences. The focus of this thesis on the math gender gap despite larger differences in the other school subjects is justified by both the correlation between children's test results in mathematics and their individual earnings perspective as well as countries' national growth correlation with a mathematically proficient labour force.

Subsection 2.1 reviews existing studies on the presence and size of a gender gap in favour of males in mathematics. Different explanations for the non-consensual findings of a math gender gap are considered. Subsection 2.2 is dedicated to three popular explanations for a performance gap: the greater male variance hypothesis, stereotyping and math anxiety. Subsection 2.3 focuses on the two-fold economic relevance of the topic. In the final subsection of 2.4, an elaboration on varying test designs and the math gender gap is discussed.

2.1 Gender differences in math performance

Although recent trends suggest that female educational attainment has surpassed that of males in many industrialised countries (Pekkarinen, 2012), empirical evidence indicates that boys continue to outperform girls in maths in most countries, even if by a small amount (Bedard & Cho, 2009; Close & Shiel, 2009; Fryer & Levitt, 2009; Hedges & Nowell, 1995; OECD, 2010).

The most recent published PISA (Programme for International Student Assessment) results from 2012 indicate that boys perform better than girls in mathematics in only 37 out of the 65 countries that participated, and girls outperform boys in five countries – their scores are non-significantly different in the remaining nations (OECD, 2014). PISA, however, is a low stakes test which could drive the finding of a limited gender gap.

According to European Commission research (Forsthuber, Horvath, & Motiejunaite, 2010), European boys and girls of the same country have similar results in mathematics at their fourth and eighth school year in most countries. Male advantage emerges only in later school years and is most relevant among students who attend the same levels of teaching programmes and year groups (Forsthuber et al., 2010).

The dataset for the statistical enquiry in this thesis is founded on the meta-analysis from Lindberg et al. (2010) encompassing 441 studies. They estimate a weighted standardized difference of 0.05 – a number in favour of males[2]. When disregarding the weight of the studies –

Box 1: Introducing Cohen's d-value

The statistical tool used to compare male and female scores across studies is the d-value. When comparing many studies, different tests are used with different scoring ranges, mean scores and sample sizes. Cohen's d-value provides a measure to standardize mean differences of male and female test score and thereby allows for comparison across studies. It is calculated per separate observation, i.e. per study in this thesis, by:

$$d = \frac{\overline{x_m} - \overline{x_f}}{s} \text{ where}$$

$$s = \sqrt{\frac{(N_m - 1)S_m^2 + (N_f - 1)S_f^2}{N_f + N_m - 2}}$$

An adjustment for the number of observations and sample standard deviations is made in this way allowing us to compare outcomes across studies.

A positive d-value indicates a larger mean score for males than females.

[2] Please see Box 1 for an introduction to the weighted standard difference, measured by Cohen's d-value.

based on the study variance and their effective - a d-value of 0.10 is found, so double the weighted size. This outcome is elaborated upon in subsection 3.2.

The article by Lindberg et al. (2010) includes another meta-analysis on post-1990 U.S. data sets, where the estimated weighted effect size was 0.07. An earlier meta-analysis by Hyde, Fennema & Lamon (1990), involves 254 studies representing the testing of 3 million individuals mainly from the U.S.. The study yielded an average d-value of -0.05, thus a small difference in favour of girls. A smaller meta-analysis from 1995 used large datasets of U.S. high school students (Hedges & Nowell, 1995). They estimated standardized differences between 0.03 and 0.26 for mathematics performance.

The results regarding gender differences in mathematics remain inconclusive, and seem consistent nor large. Some individual studies suggest a (significant) gender gap in favour of boys, but meta-analyses find at most a limited gender gap. This diversity in outcomes may be driven by varying student- and test characteristics: the different cited articles use different samples and test types and are difficult to generalize to the population at large. This is expanded upon and illustrated by additional research cited in the next two subsections.

2.1.1 Subject characteristics - nationality

Mathematics tests are made by subjects of different ages, countries, cultures and socio-economic backgrounds. The nationality of a student could be one determinant of a gender gap in performance. Depending on the country the study is held in, different gaps have been estimated. Else-Quest, Hyde, & Linn (2010) estimate the magnitude of gender differences in mathematics achievement across 69 nations. They used scores from two data sets (Trends in International Mathematics and Science Study (TIMSS) and Programme for International Student Assessment (PISA)). They found that the mean standardized difference in mathematical achievement across countries was small ($d < 0.15$). National effect sizes did vary widely however, at $d = -0.42$ to 0.40.

Fryer & Levitt (2009) in their U.S. based study consider the diversity in societal covariates of individual subjects. As controls they include girls' and boys' ratings by teachers, parental expectations and students' mother's occupation. They found none of these factors have a substantial effect on the gender gap. When testing for broader, nationwide societal features as they move to cross country comparison however, their study indicates these features are important for predicting a math gender gap. This indicates nationality may be a factor of relevance. Aspects considered are, amongst others, the level of gender inequality of the country one lives in and economic opportunities for women[3]. A positive relationship between gender equality and the relative performance of girls in maths is found for their PISA data - but not in results from TIMSS. According to Fryer & Levitt (2009), this difference could be the result of TIMSS including a large number of Middle Eastern countries that are excluded in PISA. Although these countries know a high degree of gender inequality, according to the authors', there is no gender gap in maths. The stereotype of mathematics being a male domain does not prevail there.

Also Baker & Jones (1993) found that the size of the math gender gap correlated 0.55 with the percentage of women in the workforce in those nations. Nollenberger, Rodríguez-Planas and Sevilla (2016) investigated the effect of gender-related culture, that they argue differs by nation, on the math gender gap. They analyse math test scores of second-generation immigrants to the U.S. The underlying assumption is that all subjects are exposed to a common set of host country laws and institutions. They

[3] Their overall measure for gender inequality is The World Economic Forum gender gap index.

(Nollenberger, Natalia ; Rodríguez-Planas, Núria; Sevilla, 2016)found that immigrant girls whose parents come from more gender-equal countries perform better (relative to their male counterparts) than immigrant girls whose parents come from less gender-equal countries, suggesting an impact of cultural beliefs on the math gender gap. When the authors include additional gender related factors, the transmission of cultural beliefs seems to account for at least two thirds of the overall contribution of gender-related factors' impact on the math gender gap.

The meta-analysis used as a starting point in this thesis finds that nationality was not a significant predictor of effect sizes, all effects were small or negligible (Lindberg et al., 2010). However, given that the meta-analysis only clustered by world-region rather than individual nations, this finding is not very relevant given a country impact hypothesis.

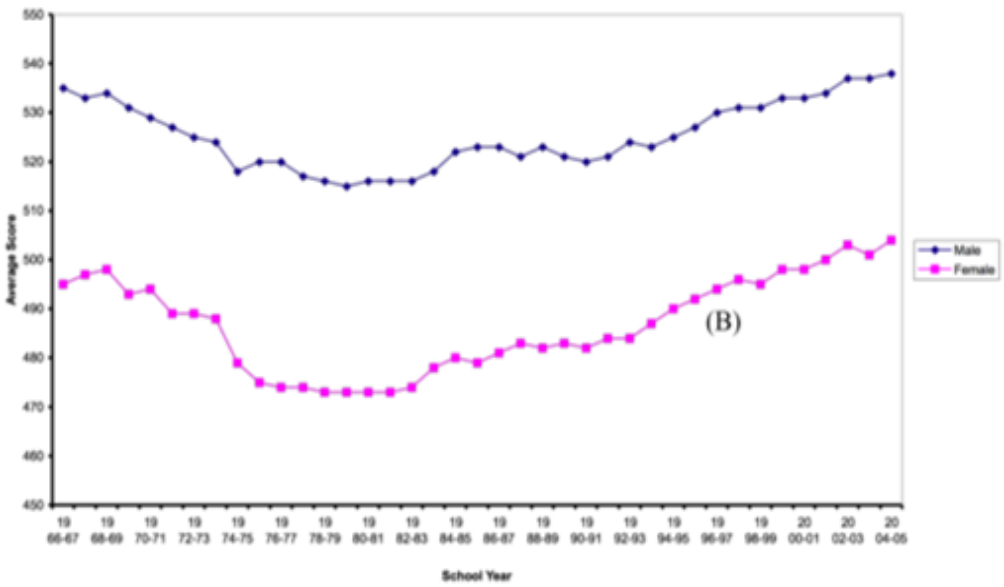
Combining the findings of studies cited in this subsection indicates that the nationality of the group of subjects in a study can impact the gender gap that is found. This is one factor that helps explain the mixed findings of a math gender gap.

2.1.2 Subject characteristics - age

The different ages of participants considerd in different studies is a second subject characteristic that supports gender gap variation. Fryer & Levitt's (2009) previously mentioned study also includes data from the Early Childhood Longitudinal Study Kindergarten Cohort (ECLS-K). They have observed that girls and boys in the U.S. scored equally high at both maths and reading skills when entering kindergarten, normally at age 5 or 6. However, by the end of fifth grade (children aged 10 or 11), girls had fallen more than 0.2 standard deviations behind their male peers in mathematics. Figure 1, taken from this study, indicates that once children reach their final year of high school and take the high-stake SAT tests, the gender gap remains profound and consistent over time. Also the meta-analysis by Lindberg et al. (2010) estimates significant impact of the age group: high school and college samples predict a d-value of +.23 respectively +.18, whereas preschool, elementary school and middle school subject groups only predicts a standardized difference of respectively -.15, .06 and -.00.

Age may thus be a driving force behind the increased performance gap that is found as subjects' ages rise.

Figure 1 Mathematics achievements on the SAT, by gender, from (Fryer & Levitt, 2009), 1966 to 2006



2.1.3 Test characteristics

Considering that school careers are based on the accumulation of knowledge, the difficulty level of mathematics that children encounter tend to rise with age. This may be a reason partially parallel to age for why the math gender gap seems to increase as schoolyears progress.

An earlier meta-analysis by Hyde et al. (1990) considers the combined effects of age groups and the depth of knowledge on test results. The math tests' items included in the dataset were coded from level 1 (simple computation: for instance, memorized math facts), to deeper understanding of concepts (level 2), or, at the highest level, complex problem solving (level 3/4). The results indicated a slight advantage for females in simple computation in elementary and middle school, and no difference in high school. There were no gender differences in understanding of concepts at any age. The more advanced levels of mathematics displayed no gender difference in elementary school and middle school. However, a gender difference favouring males emerged in high school, with a standardized difference (Cohen's d) of 0.29. This is a near tripled standardized difference as compared to the unweighted estimate of 0.10. We observe this information also in the more recent 2010 meta-analysis of Lindberg et al., whose results show that there was a small gender difference favouring male high school students on tests that included problems at Levels 3 (d=0.16), but the effect was reversed among college students (d=-0.11). As the findings are based on small number of studies, they therefore cannot be considered robust.

The decrease in the standardized difference in favour of males from high school to college may be explained by experiments being constructed amongst students majoring in the same field once in college. Boys and girls may have already opted in, or out, of certain majors or programs when entering college. Comparing gender differences when all subjects study either sociology or physics will yield smaller differences than when all students would be pooled. The majority of the college aged sample groups in our dataset are from the same field of study. This selection bias may indicate an underestimation of the gender gap in college staged settings.

In conclusion of this section: it is the complex problem solving, that is most relevant in increasing earnings and odds of entry to STEM careers (Paglin & Rufolo, 1990; Rose & Betts, 2001; Schrøter Joensen & Skyt Nielsen, 2015). None of the studies imply evidence for the thesis that girls underperform from a young age on, which would have indicated an innate inability to grasp mathematics at the level of male peers. The same hope can be expressed on the finding that different countries know different math gender gap.

2.2 Explanations for the math gender gap

The question why a gender gap exists, at least in some age groups, countries and at certain levels of difficulty, is answered by the nature versus nurture debate. Nature/biological arguments build on the perception of an innate (spatial) ability difference in favour of boys, and an larger urge for competition for males. In his influential book, "Males and Females", Hutt (1972) asserted that men are by nature better at spatial abilities, or develop them stronger due to the skills they seek to develop as children which yields their advantage in mathematics. Conform this theory, these characteristics are not susceptible to change. Education could than be used merely as a means to socialise and educate boys and girls into their respective roles as men (breadwinner, work-oriented, head of the family) and women (nurturer, carer, family-oriented).

This argumentation of an innate disadvantage has been relaxed by the test of time. For instance, over 40% of freshmen in the Bachelor's degree of technical mathematics at Delft University of Technology are women (TU Delft, 2015). Even more compelling is the said variation of study outcomes on the math gender gap across nations and age groups as discussed the previous section. This divergence suggest that strengths and weaknesses in academic subjects are not inherent, but reinforced, possibly already from a young age on. Moreover, although boys show higher means in mathematics performance, differences within the gender groups are far greater than those between the genders.

One hypothesis, that of greater male variability, seems most innate given its consistency and wide appliance and is considered in subsection 2.2.1. After that, the rest of subsection 2.2 will be dedicated to sociological and psychological theories for explaining a math gender gap.

2.2.1 Variability differences

An explanation for a math gender gap, particularly at the high performing spectrum, is the "greater male variability" hypothesis. This hypothesis could explain the underrepresentation of females in the highest math level categories. It was first posed in the 1800s and advocated by scientists such as Charles Darwin and Havelock Ellis to explain why there was an excess of men both in homes for the mentally deficient and among geniuses at the time (Shields, 1982).

The statistic used to test hypotheses of variances is the variance ratio (VR), the ratio of male to female variance in a given distribution. Variance ratios of +1.00 indicate greater male variability. Variance ratios calculated from the PISA 2003 data are shown in Table 1, and visually in Figure 2. The majority of variance ratios (from PISA 2003) is +1.00. The findings in variances is consistent though not large, with VRs ranging between 1.11 and 1.21.

The male variability theory seems also supported by the PISA 2012 summary report: "*Among girls, the greatest hurdle is in reaching the top: girls are under-represented among the highest achievers in most countries and economies, which poses a serious challenge to achieving gender parity in science,*

Figure 2: Brunner et al., (2013) based on PISA 2003 data visually

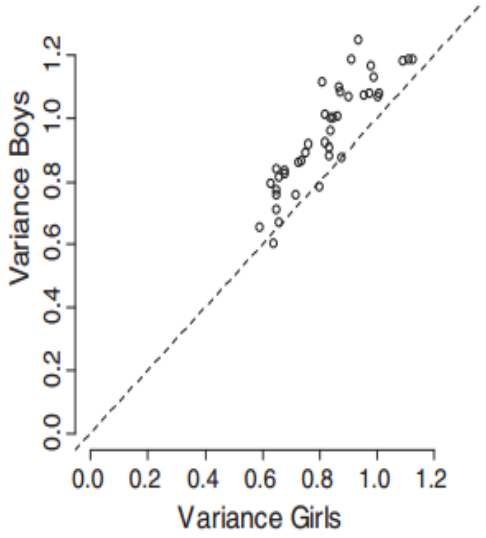


Table 1: Differences in variability in math performance between boys and girls among some selected nations. As published in Hyde & Mertz (2009)

Country	2003 PISA 15 year olds, M/F VR†	1995 TIMSS 17 year olds $(SD_M - SD_F)/SD_w$‡
Canada	1.24*	0.05
Czech Rep.	1.07	0.11
Denmark	0.99	0.01
Germany	1.12*	−0.05
Iceland	1.24*	0.04
Indonesia	0.95*	ND
Ireland	1.07	ND
Lithuania	ND	−0.06
Mexico	1.08*	ND
Netherlands	1.00	−0.13
Thailand	1.10*	ND
Tunisia	1.03	ND
Russian Fed.	1.20*	0.02
Slovenia	ND	0.01
Switzerland	1.11*	0.02
UK	1.06*	ND
USA	1.19*	0.09

*, VR significantly different from 1.0, $P < 0.05$. ND, not determined.
†Variance ratios taken from table S2 of Machin and Pekkarinen (19).
‡Calculated from data presented in table 2 of Penner (20); *P* values are not known.

technology, engineering and mathematics occupations in the future. (…) At the same time, there is evidence that in many countries and economies more boys than girls are among the lowest-performing students, and in some of these countries/economies more should be done to engage boys in mathematics."

Hedges & Nowell's (1995) widely-cited article is based on U.S. datasets comprising >150.000 adolescents and provides similar findings: boys perform better than girls by a small (though significant) amount at best. The authors argue that this may be due to the large sample size rather than a profound difference. Interestingly, the authors find substantially more boys in the high performing groups. Also Halpern & Benbow (2007) conclude the observed male advantage in mathematics is largest at the upper end of the ability distribution – our economically most relevant part. A lot of studies – also those included in the dataset used in this thesis – provide results from relatively simple tests at young ages. This is possibly one reason that in many studies only a small gender gap is found, the differences may become more pronounced only at the top levels.

2.2.2 Stereotyping

An sociological cause of the math gender gap prevailing in academic literature is the notion that mathematics and the natural sciences are stereotyped as male domains (Lindberg et al., 2010; Steffens, Jelenec, & Noack, 2010).

A persisting stereotypes of women's limited mathematical ability could have short and long term effects in two ways. First, girls having lower mathematics results because they believe that they are inherently inept and self-sorting into lower tracks. Second, this (unintentionally) discourages women from entering or persisting in careers in the STEM subjects. Not merely because of a lack of self-esteem but rather lack of role models and nudging. Stereotypes have a harmful effect on actual performance and job opportunities alike, as argued and tested in studies from respectively Giulia, Silvia, Francesca, & Caterina (2014) and Ernesto, Paola, & Luigi (2014). Moreover, cognitive social learning theory implies that stereotypes can influence self-efficacy or competency beliefs: lower expectations for girls' performance in maths compared to boys do exist, regardless of the decreasing gap over time in many nations and many individual girls with strong mathematics skills (Bandura, 1994; Else-quest et al., 2010; Lindberg, Hyde, & Hirsch, 2008; Tiedemann, 2000).

Maccoby (1966, p. 40) describes the longer term effects of gender dominated schooling domains as: *"Members of each sex are encouraged in, and become interested in and proficient at the kinds of tasks that are most relevant to the roles they fill currently or are expected to fill in the future".* Hence, skills one does not expect to need later in life will influence one's investment in – and hence attainment of – such skills.

Research by Pope & Sydnor (2010) indicates that such stereotype variation already occurs at the state level in the United States. They find male-female ratios of students scoring in high ranges of standardized tests vary significantly across the US. The resulting variation is systematic in several ways. First, states where males are highly overrepresented in the top math and science score ranks, also tend to be the states where women are highly overrepresented in the top reading scores. This pattern suggests that states vary in their adherence to stereotypical gender performance, rather than favouring one sex over the other across all subjects (Pope & Sydnor, 2010). Second, the genetic distinction and the hormonal differences between sexes that might affect early cognitive development

(that is, innate abilities) are likely the same for both sexes regardless of the state in which a person happens to be born. This suggests environments significantly impact gender disparities in test scores.

A questionnaire amongst 11,500 women between the ages of 11 and 30 in 12 European countries by Microsoft indicated that girls became interested in STEM at the age of 11-and-a-half, but this starts to wane by the age of 15. As a key reason to not follow a career in STEM, girls cited a lack of female role models (Microsoft, 2016). OECD researched data (stemming from 2006) that point to similar results: Appendix 8.1 shows more elaborate outcomes of the PISA 2006 study on the perception of school subject importance of the genders in OECD countries, and indicates gender differences in line with the stereotype. The level of stereotyping of STEM as a male domain remains however undecided: according to the earlier 2003 PISA results boys and girls are similarly interested in the natural sciences; and there is no overall difference in boys' and girls' inclination to use science in future studies or jobs (OECD, 2004).

2.2.2.1 Overcoming stereotype threat – strategic marketing

If the goal is to draw a larger work force into the STEM fields by including women, one aspect to overcome is the masculine image of STEM in general and mathematics in particular. This as cultural stereotypes about gender have an impact on students' career aspirations and subject choices (Bedard & Cho, 2009; Correll, 2001, 2004). Research suggests that subtle nudging may help to further increase girls' interest in the STEM field.

Harvard professor in economics Claudia Goldin, for instance, underscores framing in the light of female interests. She proposes that women may (by nature or by nurture) value supporting other people in society more than men do. Portraying STEM jobs in a way corresponding to this may nudge women to enter such fields of profession and work on their math skills already at a younger age (as recorded by Dubner, 2016a). Also Johnson (2007) describes other non-skill related barriers to female science-interested youngsters in their continuation in STEM. He points to a lack of sensitivity to their difference, discouragement, and a sense of alienation from school science. In describing the experience of these women moving through undergraduate science, Johnson (2007) concludes: "*The first step in making science more encouraging (…) is for scientists to recognize that science has a culture, and that certain types of students may find it challenging to understand and navigate this culture.*" (p. 819). If you raise students with the idea that all can be taught and there is no innate lack in potential, the context may be framed in a way that stimulates a decrease in the math gender gap.

Scott, Page, & West (2010) underscore the importance of female role models. They researched the impact of a female professor in the exact courses for U.S. Air Force Academy students. Their results suggest that professor gender has little impact on male students, but does have a powerful effect on female students in terms of performance and STEM continuation. The estimates are largest for students whose SAT math scores are in the top 5% of the national distribution. Of note is that the gender gap in course grades and STEM majors is eradicated when high-performing female students are assigned to female professors in mandatory introductory math and science coursework.

As long as (unintended) stereotypes push individuals from either gender into a certain direction – a person is apt to be influenced and develop a set of skills in line with what they expect to need later in life.

2.2.3 Math anxiety

Despite performing equally well as boys in most countries, girls tend to have a weaker self-concept in the sciences than males. This is the third factor I would like to look into after greater male variability and stereotyping that can help explain a math gender gap and causes girls to perform worse at certain times and levels (Forsthuber et al., 2010). This higher level of anxiety and lower self-esteem when it comes to (applied) mathematics is a wide cited cause of the math gender gap. When experiencing distress when performing a certain task (such as math), evidence suggests this may harm your results.

One may be sceptic given the potential of a two-way relationship: if one actually performs worse it make sense to be more anxious about it. From quoting the 2012 PISA report we learn the following: "PISA results show that even when girls perform as well as boys in mathematics, they tend to report less perseverance, less openness to problem solving, less intrinsic and instrumental motivation to learn mathematics, less self-belief in their ability to learn mathematics and more anxiety about mathematics than boys, on average; they are also more likely than boys to attribute failure in mathematics to themselves rather than to external factors." (OECD, 2014). Also Cheema & Galluzzo (2013) find, using U.S. PISA 2003 data, that both anxiety and self-efficacy contribute significantly towards explaining variation in math achievement. Moreover, their results show the gender gap disappears once important predictors of math achievement, such as math-specific self-efficacy and anxiety, are controlled for.

An OECD research found that on average, boys had higher self-efficacy, i.e. a higher level of confidence, in tackling specific math tasks. Boys also had higher levels of belief in their mathematic abilities than girls. Girls had higher anxiety levels regarding mathematics - although both genders are subjected to math anxiety. More than 60 per cent of 15-year-old females and half of the males report that they often worry that they will find mathematics classes difficult and that they will get poor marks. For other subjects no anxiety levels are tested. Poland was the only country showing no significant gender difference in either levels of self-efficacy nor in self-concept and anxiety in mathematics. Italy showed no significant gender differences regarding self-concept and anxiety (OECD, 2004).

The male lack of anxiety may coincide with a larger willingness to compete. Buser, Niederle, & Oosterbeek (2012) find that gender differences in competitiveness account for a substantial portion of the gender difference in academic track choice, conditional on ability. If one feels more prone to a challenge and is self-secure about potential accomplishments, one may easier enrol in more advanced classes and develop skills accordingly, this lack in females could enlarge a math gender gap particularly as higher levels become available. In mathematics and science there is, potentially due to this, a tendency of females to participate in different higher-level school programmes or streams than their male counterparts. Research by Forgasz (2006), Helme & Lamb (2007) and Watt, (2005) indicates that there are more female students choosing lower-level math courses than males, and that this difference is not based upon mathematics achievement. This suggests it is not ability that drives gender differences in higher level math course enrolment and implies there are other factors of relevance than a prior achievement gap between the genders.

In conclusion of this section: different lines of reasoning are available but no single theorem is preferred widely to account for a math gender gap. There is likely an innate advantage for males in the top ranks that drive part of a gap – at least in top-levels. Such advantage may also stem from innate or taught differences in self-esteem and urges for competition. It is particularly the factors of stereotyping of math as a male domain and math anxiety amongst female students that may lower their enrolment

in (more advanced) math courses, and limits them from reaching certain levels. Controlling for prior performance, female students are over-represented in lower-level math courses and under-represented in higher-level math courses. I am reserved to state this is for anxiety choices alone: girls may seek and find value in other courses and developing other skills.

What is of interest in this particular thesis is the possibility that test anxiety diminishes as time progresses, so the longer someone has, the more time one would have to calm down and ensure the knowledge needed is mentally accessed. If this is the way in which test length impacts the gender gap, rather than speaking of female favoured bias, it could also diminish male favoured bias – not of design but of a gender's tendency. This role of math skills in the payment gender gap and nationwide reliance on a STEM educated labour force will be considered in coming subsection 2.3.

2.3 Economic relevance of a math gender gap: math income correlation and nation economy's STEM field stimulation

There are two economic aspects that require a just assessment of mathematical skills: the skills' correlation with individual income and its relation to national economy and innovation. First, a level academic playing field for both genders is fair due to the positive correlations between math skills/math course taking and income levels. Keeping track of actual performance (and performance differences) requires an assessment that does not favour some groups over others. Second, the wish for a STEM focused labour force is at times seen in nations' policy that aim for innovation and economic growth (Beede & Julian, 2011; Hanushek & Woessmann, 2008). Such policy is understandable given that a larger share of the population educated in the STEM fields corresponds with national and technological innovation (Hanushek & Woessmann, 2008). The link with a math gender gap in math performance at the higher levels is relevant in this as potential female STEM employees not joining these sectors are a loss to the workforce. In the coming two subsections these two economic relationships of mathematics ability are considered in more detail.

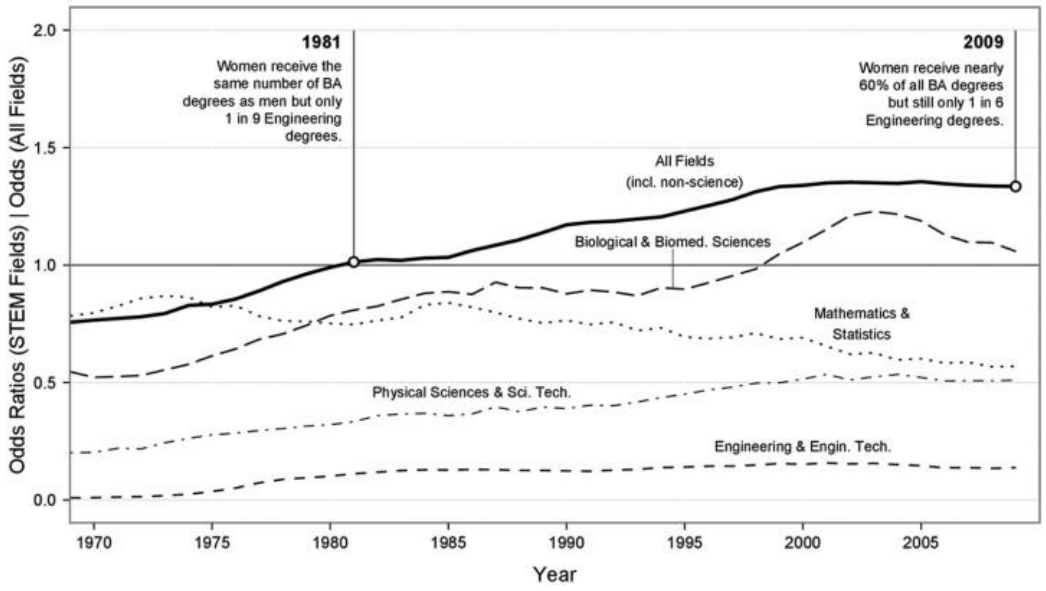
2.3.1 Relation to individual labour income

The study levels most relevant to earnings are the higher math levels, in the high school and beyond phase (Schrøter Joensen & Skyt Nielsen, 2015). Rose & Betts (2001) address the hypothesis of the impact of math skills on labor income: they use an instrumental variable and find that completing more advanced math courses has a larger impact than completing less advanced courses on both labor market outcomes as well as on bachelor's degree completion. The importance of mathematical skills for income has also been documented by Joensen & Nielsen (2010). Their findings indicate that math skills have a causal effect on labour market outcomes. There is evidence that the individual returns to maths skills are higher than the returns to other skills (Buonanno & Pozzoli, 2009; Grogger & Eide, 1995; Koedel & Tyhurst, 2012; Paglin & Rufolo, 1990). Note that there is a pipeline effect from one math level to the next as before learning advanced skills one starts with the basics. When (a group of) children are discouraged at an earlier stage and do not select or enrol into the higher levels, the higher math levels will not be accomplished by them.

Let us consider the trend of individual math skills attainment in the U.S. and Europe at the higher level. For the U.S. we see the increase in women pursuing careers in STEM is uneven across the different fields. In 1970 only 14% of the U.S. doctoral degrees in the biological sciences went to women, in 2006 this figure had risen to 49% (Snyder, Dillow, & Hoffman, 2008). Entry into other STEM areas has been

slower. Figure 3 visualizes this U.S. ratio of females to males for receiving a bachelor diploma in the different science fields. The ratio for the fields of mathematics and statistics is slightly decreasing and the low rates of women in engineering and engineering technology are constant. This is particularly noteworthy in comparison to the steady rise of women into overall tertiary education (visualised by the 'All Fields' line).

Figure 3: Ratio of female to male bachelor's degrees awarded by field of study, 1969-2009. Source: Digest of Educational Statistics (2009: Tables 268, 299, 303, 305, 312 and 313). Note: The trend line for all fields shows the odds that any BA degree is awarded to a woman, and the lines for the different subfields show the female/male odds ratio for the respective STEM field. As published in Legewie & DiPrete (2014).



2009 ACS data on undergraduate fields of study in the U.S. show that women account for nearly half of employed college graduates age 25 and over, but only about 25 percent of employed STEM degree holders and an even smaller share – about 20 percent – of STEM degree holders working in STEM jobs (Beede & Julian, 2011). While both males and females attain equal numbers of bachelor's degrees in math and statistics, in fields where mathematical preparation and application plays an indirect role - such as in science, technology, and engineering - women lag behind (National Science Foundation, 2016).

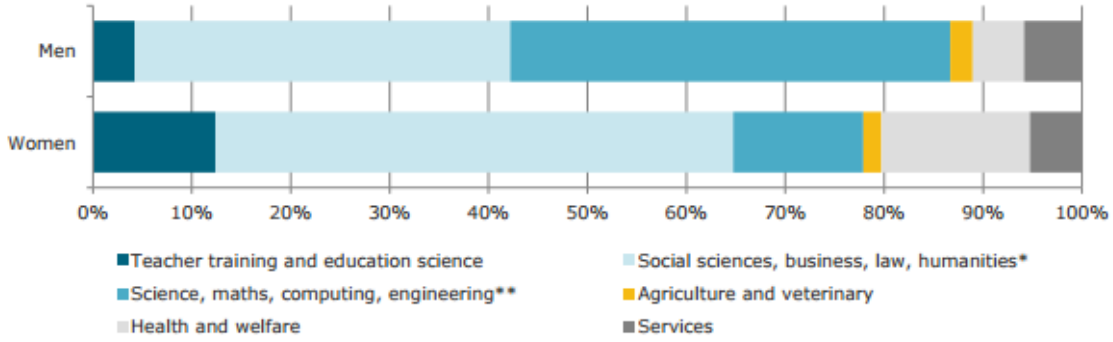
This can be seen in perspective to the European developments. Eurostat data shows the female/male attendance in tertiary education in STEM fields grew by 7 percentage points from 2006 to 2015, see Table 2 – with a +50% increase in students of both genders. The jobs in STEM (possibly due to a lag in the move from education to jobs) show a smaller absolute increase for both genders, where we see a shift from female domination of the field to male domination.

Table 2: based on Eurostat data (values x1,000)

		2006	*2015*	*Increase 2006-2015*
Followed tertiary education in STEM	Female	4,014	6,553	63%
	Male	4,507	6,793	51%
	Female/Male	0.89	0.96	7 percentagepoint increase in favour of females
Job in STEM	Female	2,313	2,596	12%
	Male	2,157	2,757	29%
	Female/Male	1.07	0.94	13 percentagepoint increase in favour of males

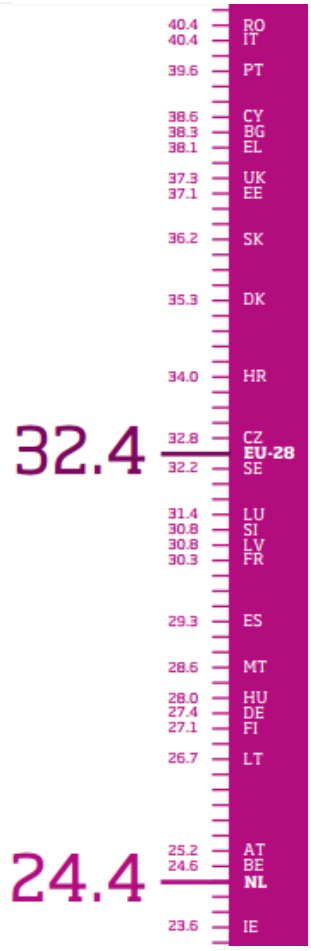
Figure 4 provides a more detailed picture of EU study field distribution, although the rise in absolute numbers of students is not apparent from the picture (European Commission, 2015). At glance, this seems comparable to the pattern in the United States. Figure 5 indicates the share of women in the European Union with a degree in science, technology, engineering and maths (the STEM fields) per country. According to the CBS publication the 32 percent share of female students in STEM has been constant in the last ten years (CBS, 2016).

*Figure 4 Distribution of study field for men and women. Source: Eurostat (LFS,2014), based on a May 2015 extraction. the indicator shows tertiary education attainment amongst 30 to 40 year olds by fields of study; general programmes and unknown field of study. *=also including languages and art, **=also including manufacturing and construction.*



The gateway to many high-paying STEM jobs (but also high-paying non-STEM jobs) is a STEM degree. The line of argument of a positive link between improving mathematics skills and better labour market perspectives is at times disputed. Schooling in STEM fields may rather signal foundational knowledge or ease of learning developing mathematical skills and for that reason yield higher earnings. Weinberger, moreover, found that only 1/3 of the college-educated white U.S. males in the STEM workforce had high school quantitative SAT scores of +650 (as cited by Hill, Corbett, & Rose, 2010). Progress in STEM fields is fuelled not only by those highly talented, but also by the millions of laboratory technicians and other bachelors- and masters-level scientists whose mathematics skills might place them below the 75th percentile but whose contributions are still essential. From a gender equality perspective this is relevant due to the variance differences as elaborated on in section 2.2.1. This larger variability may hence be of limited relevance from the perspective of an equal economic opportunity.

Figure 5: percentage of girls in STEM education EU-28, 2012-2013



Noteworthy in this respect is a study by Bastalich & Mills (2003). They conducted a qualitative study of female and male experiences in a range of engineering disciplines, industry sectors and work locations. They identified a significant contributor to reasons for women leaving the profession (so despite the right education) to be: 1) a feeling of alienation within the prevailing workplace culture and 2) this more so than family responsibilities, or lack of confidence, technical expertise, or interest in engineering work compared to men. This study hence

supports the notion of STEM fields being a male dominated domain, with its consequences for female participation. These findings however do not diminish the relevance of keeping unbiased track of math scores of all students, regardless of gender and stimulating the learnings of the STEM fields

2.3.2 The role of math, STEM and innovation in economic growth

The second economic relevance of math comes from findings of amongst others Hanushek & Woessmann (2008). Their research indicates that a larger share of the population educated in the STEM fields (science, technology, engineering and mathematics - all fields related to mathematics) corresponds with national and technological innovation. According to (semi-)endogenous growth theories as developed by a.o. Grossman & Helpman (1991) and Romer (1990), the stock of science and engineering graduates is an important determinant of the innovative capacity of a country, because the supply of R&D workers determines the total amount of R&D activities which can be carried out – and STEM students are arguably well-represented in R&D. R&D in turn generates technological change and increases in economic growth or income per capita. From a national perspective, investing in and stimulating mathematic skill development could therefore be justified. Such investments are indeed made: in 2015 the U.S. set up a committee on STEM education and a five year plan to facilitate education in STEM from the preschool through graduate education expecting growth (U.S. Department of Education, 2015). Before this, the U.S. President's *Council of Advisors on Science and Technology Report* targeted postsecondary STEM education strategies to increase the number of STEM graduates by one million students over the next decade (as explained in Olson & Riordan, 2012).

The European Union set up the EU STEM coalition who set as a goal to 'raise awareness among governments, industry and education about the crucial role of STEM in our society'. Sensible, as a European Committee report dedicated to the topic of the STEM labour market indicates lower unemployment rates in STEM as well as labour market shortages in these fields (Caprile, Palmén, Sanz, & Dente, 2015). A study, though published over ten years ago, by Noailly, Waagmeester, Jacobs, Rensman, & Webbink, (2005) investigated the Dutch situation. The authors do not find evidence for scarcity in the labour market of these graduates, it even seemed to have weakened. They point towards the potential increased internationalisation of the science & technology labour market to explain the shortages that employers expressed. The EC report highlights the imbalance of the genders and the opportunities in this respect (Caprile et al., 2015). Wealthy countries that fail to make the STEM work field attractive for as many citizens as possible are at risk of producing too few citizens with the skills necessary to compete in a knowledge-based economy driven by science and technology. Hence, including women and men into pipelines to these careers and tracking their performance in an unbiased manner is relevant.

When looking into this labour market shortage one indication of a 'leaky pipeline' could be apparent at the start of the labour market. In the United Kingdom a questionnaire amongst STEM-students was held to look into this. It found the vast majority of final-year students, at undergraduate through PhD level, report that they do want to pursue a career related to their degree subject, although some were more definite about this than others and this proportion varies somewhat with degree subject. The most likely reason students seek employment in a direction away from STEM is because other fields are seen to be of more interest, although more practical and career-related reasons are also significant for graduates considering 'leaving STEM'. The profile and reputation of certain major employers, especially in STEM Generalist and non-STEM sectors, with well-established and substantial graduate

schemes, were attractive and powerful influences on 'undecided' graduates at the transition stage between university and work (Mellors-Bourne, Connor, & Jackson, 2011).

In conclusion of this subsection: math is important via two ways. Obtaining math skills and staying in the pipeline of a STEM education is for individuals of value from a labour income perspective. These skills are valued by the market and given a gender gap in STEM degree obtainment, an unbiased math assessment is of value. Second, for a nation a labour force skilled in math has more innovation potential via the STEM fields also.

2.4 Importance of test design & performance decline

The literature review thus far elaborated on the economic relevance of, and the diversity in findings and reasons for, the math gender gap. The remainder of this section is dedicated to test format, and determines whether this could also serve as a driving force behind mixed findings. This thesis' main aspect of interest of test format is the length, which will be elaborated on in the last subsection of 2.4. Before that, other aspects of format are considered.

The working assumption in this paper will be that in general tests can serve as a reasonable indicator of someone's skills in, in this case, mathematics. The provision of up-to-date unbiased information on mathematics performance is important both for individual students and the skilled labour force. However, when measuring ability in a test, and its design favours one gender over the other, this implies a bias and could partially explain the differences in outcomes of studies investigating the math gender gap. There is evidence on the existence of gender differences in testing behaviour independent on knowledge or ability on the evaluated topic. Awareness of this may be the first step to overcome such bias. This section considers three researched possibilities of bias (related to risk aversion (2.4.1), competition (2.4.2) and answer formats (2.4.3)) and the hypothesis of this thesis: test length (2.4.4).

2.4.1 Disadvantage of the doubt

The first example is the 'disadvantage of the doubt' that may occur on exams. In an experimental setting Baldiga (2013) found that girls have a significantly lower willingness to guess in multiple choice tests, possibly due to a higher risk aversion, as was discussed in subsection 2.2.3. Espinosa & Gardeazabal (2013), Pekkarinen (2015) and Tannenbaum (2012) use data from a field experiment: if there is a penalty for making a mistake, more women than men tend to leave answers blank – and they miss out on points. In the high-stakes SAT test women prefer not to fill in an answer, rather than guess. Coffman found that if the penalty for answering is taken away, the gender gap in score also went away (as cited in Dubner, 2016b). As the expected value of making a guess is at times rewarded in tests, this format negatively affects girls' scores.

2.4.2 Competitive environments

Second, in competitive environments women of all ages seem to score lower. Girls tend to shy away from competing, particularly if their results will be compared to boys' – which indivertibly happens in maths predominantly on the more advanced levels given the limited female representation. There is a large range of articles published documenting gender differences in performance under competitive environments, see Gneezy, Niederle, & Rustichini, (2003) and Gneezy & Rustichini, (2004) among others. Azmat, Calsamiglia, & Iriberri (2016) indicate that girls perform worse compared to boys as the stakes of a test increase. As the competitiveness of an environment increases, the performance and participation of men increases relative to women (Croson & Gneezy, 2009).

Ors, Palomino, & Peyrache (2013) find that males obtain higher test scores when they are competing for college seats than predicted by their previous grades, while the opposite is true for females – they score worse. Also a Dutch research by Buser et al. (2014) confirms this. They examine the predictive power of the later important choice of academic track of secondary school students. Even though boys and girls display similar levels of academic ability, boys choose substantially more prestigious academic tracks, where more prestigious tracks are more math- and science-intensive. Their experimental measure shows that girl are also substantially less competitive than boys. Buser et al. (2014) find that the gender difference in competitiveness accounts for a substantial portion (about 20%) of the gender difference in track choice.

Ergo: capable female candidates may not live up to their potential particularly at important times. Vice versa, the motivation of girls to perform regardless of test importance may underestimate male ability prior to the test. A cautious note on these studies is that high-stake exams may unintentionally be designed in favour of boys due to their multiple choice format (elaborated on in next section). The design may be a stronger driving force beyond mere competition when seeking to explain the gender's difference in performance.

2.4.3 Answer formats

An open answering item format may correlate with an advantage for females, whereas multiple choice formats is generally considered advantageous for boys (Beller & Gafni, 2000; Reardon, Fahle, Kalogrides, Podolsky, & Zarate, 2016). Beller & Gafni (2000) investigated the 1988 and 1991 results from the International Assessment on Educational Progress mathematics test measured in six countries. In the 1988 assessments they find that gender effects were larger on multiple choice than open exam items. However, the 1991 assessment produced contrary results: gender effects tended to be larger for open exam items than for multiple-choice items. Further investigation of the data revealed that the inconsistent patterns of gender effects were related to the difficulty level of the items, regardless of item format. Correlations between item difficulty and item gender effect size were computed for the students (age 13) in the 1988 assessment and for the ages 9 and 13 in the 1991 assessment. The correlations obtained were 0.26, 0.47, and 0.53, respectively, suggesting that the more difficult the items, the better boys perform relative to girls. These findings match up with the elaboration of section 2.1.3., where depth of knowledge is argued as a driving force of the different gender gap outcomes across studies.

Reardon, Fahle, Kalogrides, Podolsky, & Zarate (2016) exploit the differences in school districts' results from nationwide tests as compared to district specific tests in the United States. This is possible as district specific tests vary substantially in the proportion of multiple-choice items on their tests in mathematics (a range from 50-100% multiple-choice). Their findings reveal that boys do better on multiple-choice tests than girls of the same academic skill. These results appear to be driven primarily by gender-by-item format interactions affecting performance. On mathematics tests, the difference in performance (favouring boys) is roughly 0.20 to 0.30 SD larger in multiple-choice tests than on constructed response item tests, favouring girls less and boys more on multiple-choice tests than on constructed-response tests. These patterns are consistent regardless of whether nationwide NAEP (all 50 states) or NWEA (Northwest Evaluation Association, including 3,700 school districts) tests are used as the audit test.

2.4.4 The impact of timing and test length

As a final factor of potential test design bias, and the central bias under research of this thesis, I look into the limitedly researched test length. We here consider some research in chronological order. Zoller & Ben-Chaim in 1989 looked at the relationship between anxiety, achievement and test design of college science students in Israel. They found that females significantly more than males prefer the 'take-home' exam where any supporting material can be used and time restrictions are virtually unlimited. A written exam with no time limit and access to any supporting material was popular with both genders. Written exams with no supporting material allowed and under a time restriction were the least desirable type for students of both genders. Zoller & Ben-Chaim (1989) also found that the test anxiety of female students was higher than that of males in traditional written exam conditions but dropped substantially in take-home examinations – an indicator for different levels of achievement under more time pressure. In research by de Lange (1987) boys did better under time pressure than girls (de Lange, 1987). The length, however – was not considered by de Lange.

Hannabus (1991, 1992) counters McCrum's (1991) statement that women were disadvantaged by the examination system in Oxford at the time of writing. He looked into the situation and found that until 1976 the exam performance of males and females was remarkably alike – regardless of the standard used 'male preferred' multiple-choice format. He finds the evidence concerning time limits of the divergence in performance between male and female students more convening. One could argue that at that Oxford was only attainable for the most talented females of the population, indicating selection bias.Mehrens, Millman, & Sackett (1994) obtained results that indicated different results. Non-disabled (so ordinary) test-takers on the Multistate Bar Examination significantly improved their scores with extended time. In a test among abled (and disabled) students no gender interactions were found for either the verbal or math section when the test was extended 1,5 or 2 times. Such an outcome contradicts the hypothesis of gender difference in test length preference.

The unpublished study by Balart & Oosterveen (2016) looks into an indication of a decreasing gender gap as the test end nears and sparked interest for the current tehsis. Their findings are based on the PISA test. Balart & Oosterveen (2017) find performance decline to be larger for boys than girls in 71 out of the 74 countries and this is statistically significant for 58 of them (64 if a one tailed test in the positive direction is used) on a 10% significance level. Using a non-linear Wald test, statistical significance is found for 67 countries at 10% significance level (64% for a significance level of 5%). This indicates that performance decline is a potential bias threat for tests as an evaluation mechanism. For most of the PISA participating countries, the gender gap in the questions on reading exacerbates as long as the test goes on whereas differences in science and mathematics shrink. The authors suggest that therefore shorter tests could reveal a larger gender math gap. One should, however, consider that the PISA test used in the analysis from Balart & Oosterveen (2016) is a low-stakes test. Under a situation with less competition - where the gender gap also tends to be smaller – boys and girls may respond with different performance decline than in high stakes test. The current analysis aims to widen its scope by including studies with different lengths and stakes.

In the literature review I have discussed the diverse outcomes of studies and potential reasons for this diversity. The economic factors that make math assessment a relevant topic in terms of equity and economic opportunity have been considered and the test design as one factor of potential assessment bias. This hypothesis will be the basis for the remainder of this thesis, which is dedicated to assessing the influence of test length on the math gender gap by using statistical analysis.

3 Data collection

To assess the hypothesis of female favoured bias as math tests lengthen, statistical analysis is used. To make use of the existing data I expanded a meta-analysis of peer-reviewed articles with measures of test length. The final dataset is an extended version of that used by Lindberg et al. (2010) where two measures for test length are added: the number of questions a test contains and the maximum number of minutes allowed to finish a test.

3.1 Data collection

In 2008, Lindberg et al. decided on the following method to identify the studies of interest for their meta-analysis:

> "Computerized database searches of ERIC, PsycINFO, and Web of Knowledge were used to generate a pool of potential articles. To identify all articles that investigated mathematics performance, the following search terms were used: (math* or calculus or algebra or geometry) AND (performance or achievement or ability) NOT (mathematical model). (…) Search limits restricted the results to articles that discussed research with human populations and that were published in English between 1990 and 2007. The three database searches identified 10,816, 9,577, and 18,244 studies, respectively, which were considered for inclusion. (…) 3,941 studies met the aforementioned criteria. These articles were then printed and examined to determine whether they presented sufficient statistics for an effect size calculation. The final sample of studies included in study 1[4] utilized data from 242 articles, comprising 441 samples and 1,286,350 people."

For further details on collecting the dataset and grouping the studies, please examine the original publication's *Method* section (Lindberg et al., 2010).

The original dataset of the article was imported to StataSE 14 (64-bit version). The observations in the dataset were matched to their original published study by linking the unbiased effect size and raw variance ratios: information that was available both in the data as in the online appendix. By examining the articles and making information requests, the two new variables on test-length could be added. In Appendix 7.2 one can find a complete overview of the articles that make up the dataset. Note that one article could contain multiple studies. Also added to the dataset were country ISO-code, website links to original articles, contact details of authors or relevant institutes and potential problems for statistical analysis of the observations. The existing variable on whether tests were timed or not was verified and often added or updated.

The original articles of the observations were retrieved via the Erasmus University network and partially from other Dutch universities' networks. Not all information could be retrieved from the articles directly, so authors, ministries, test publishers and educational institutes were contacted via e-mail for insights on test length. Although the final dataset includes the lengths of many tests, there remain missing variables when a conclusive length for observations was not obtainable. For instance because contact was not possible, authors were unable to recall the lengths, or multiple tests with different lengths matched with a single observation in our dataset.

[4] The 2010 article by Lindberg et al. consisted of two meta-analyses. 'Study 1' refers to their meta-analysis on the 441 samples, which will be explored in this thesis. 'Study 2' is a meta-analysis based on fewer but larger studies and is not further researched in this thesis.

3.2 Summary of findings from Lindberg, Hyde, Petersen and Linn (2010)

As a measure for gender difference, the standardized mean value of the difference (Cohen's d, also known as raw effect size) is used as it enables one to compare findings across studies. It is a measure of the distance between the male and female means in standard deviation units. A positive d-value indicates superior performance of boys. There seems no academic consensus on what entails a small or a large d-value.

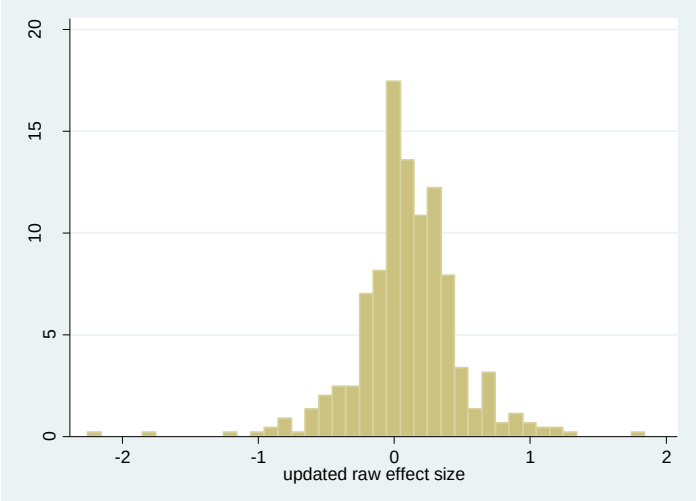
Lindberg et al. (2010) find an overall weighted d-value of 0.05. When not accounting for these inverse variance weights the mean d-value of all studies doubles to 0.10. An explanation of inverse variance weights is found at the start of section 4. A reason for this doubling is found in the study with the largest sample size. This observation relates to the study by Held, Alderton, Foley, & Segall (1993) and is based on data from a U.S. navy application test with a female to male ratio of 1:6. The d-value of this article was -0.29, indicating a considerable female advantage in math performance. However, in this study women applied for other positions than men, for which a higher educational backgrounds was required. Hence, their higher math scores are of no surprise given their higher previous education. - there is selection bias. Given that the focus of this article was not necessarily on the math gender gap, this occurrence is legit but harmful to our dataset. A remark related to this topic is made in the article itself:

"*A high school diploma requirement for female but not male enlistment may have resulted in a restricted range of female talent due to higher academic ability*." (Held, Alderton, Foley, & Segall, 1993)

Appendix 8.4 visualizes the impact of including this observation on the estimated weighted standardized gender difference. When excluding this observation but using weights, I find a standardized difference of 0.14 – so nearly triple the 0.05 weighted value noted by Lindberg et al. (2010) and 40% higher than the unweighted average effect size. The weighted outcomes excluding this study are considered our regression of most value in the analysis in section 5.1.

When we look at all the observations of the meta-analysis we learn the standardized effect size ranges between -2.3 to 1.8 across the studies. At times there is a math gender gap in favour of men, at times in favour of women. Figure 6 graphically represents the largely symmetric distribution of the raw effect sizes of the studies. There are 31 outliers – evenly spread below and above zero (1.5 IQR away from

Figure 6 Raw non-weighted effect sizes, as based on study by(Lindberg et al., 2010)



the 25% or 75% point). Excluding these outliers does not alter the estimated unweighted raw effect size.

The standardized differences had to be re-calculated due to minor alterations in the dataset – as will be elaborated on in the coming subsection 3.3. The original values differ slightly from the updated, also in part due to rounding differences as Stata was used for this paper's analysis rather than SPSS as used by Lindberg et al. (2010). The correlation between the original and updated effect size remains high: 0.9912. This can be regarded a sufficiently solid basis to build the analysis upon – where the updated values are used.

As the focus of this paper is not on the gender gap itself but its relation to test length – I would like to refer to the tables in Appendix 8.5 for a further elaboration of the original meta-analysis' findings.

3.3 Extending the dataset with length measures

The originally entered data largely matched the original articles. Some adjustments were nevertheless required as noted from re-reading articles. To trace these changes, the merged dataset includes additional variables whose names end with _A (first letter of my name). The original variable names end with _L (from Lindberg) in the final datafile. The updated values ending on _A are used in this data analysis.

The following alterations were made for the final dataset: the variable indicating whether a test was timed was occasionally adjusted: 125 observations remained identical, twelve were adjusted and 155 were added in the final version. Three male sample sizes were adjusted in the final dataset. This was the case for four observed sample sizes of females. Six mean value scores of males were adjusted on the basis of articles and six female mean values. Regarding standard deviations, updated values were included for three studies for male and four for female observations. Six observations were put in the wrong world region, so these were also updated. In general, errors seem to be made due to typo's (e.g. 32 instead of the actual 23 sample size would be noted) or mixing up male and female values.

From re-reading the original articles it appeared potential issues for the research were in order. These will be accounted for in the data analysis section when considering sensitivity in subsection 5.1, and are the following:

a. One double observation (once).
b. One observation untraceable to an original article – and no article left to assign it to.
c. 20 articles could not be found online or could not be accessed – this corresponded to 32 observations.
d. For 25 observations the original data input was not incorrect but data that could be of added value in our research was not incorporated as an observation. For instance when an average of multiple tests with different test lengths was included as one observation, rather than noting these as multiple observations. This is also explained in the Lindberg et al. (2010) article itself for the cases where multiple effect sizes were available for the same sample (see p. 1126). Hence, some information that could have been used was not, this may be relevant when using this dataset in future.
e. 38 times schoolyear grades were used or the composition of the math measure was hard to retrace. For instance there was some variance in time for different participants and these findings were merged in the dataset. Also, at times, d-values would be available in the dataset without means or standard deviations provided in the original article. Although the primary

dataset authors may well have accounted for this for instance by directly requesting this information at the authors, we will exclude these cases in subsection 5.1.

f. For 49 studies the author did not reply, died or retired, hence no updated information was available for these studies.

g. 33 times the test duration was an estimate. This happened when the original study author indicated this her- or himself, e.g. she would be sure the test was 30-40 questions long, so I took an estimated test length of 35 minutes for our dataset. Those studies where expected test length was retrieved from secondary sources, or retrieved by adding the maximum duration of verbal math exams where there would be x seconds per question before the student would have to move on with the next questions, were also labelled as estimates.

In sum, there were six observations (representing 8.352 subjects) excluded at all times from the data analysis. Once because the observation was identical to another (problem a), once because it could not be retraced what original study it belonged to (problem b) and four times because the value did not correspond to mathematical ability, but to more ambiguous measurements. For instance scores based on self-assessment on problem solving technique, percentages of children in certain categories or where literacy ability was half of the test. This was probably overlooked or counterargued in the private reasoning of Lindberg et al. (2010) but I preferred not to take the risk of including them in my analyses. The mean raw effect size value of those six excluded was 0.069 (0.12 se), which is not significantly different from the unweighted 0.10 raw effect size. This was checked for by a two-sample t-test.

3.4 Description of statistics

The key summary statistics of the final dataset are presented in table 3. In Appendix 8.2 a more extensive data summary of variables is provided. For most values we see considerable standard errors, indicating wide ranges. The d-value (raw effect size) varies, with the distribution spread evenly around a value of .10 (as could be seen from figure 6 in section 3.2.). The 435 studies of which total sample size n was known, were primarily small samples: only 75 studies had a sample size above 1,000 participants. The number of studies by amount of participants are plotted in two separate figures for easier graphical representation, namely Figure 7 and Figure 8. We see an even spread amongst the four mid age groups and few observations from pre-school and general population age groups, visualized in Figure 9. As can be seen from Figure 10, the majority of studies' have participants of general ability.

Table 3 Summary statistics

Variable	Observations	Mean & (std.err.)	Min	Max	Visualisation
D (raw effect size)	441	.10 (.38)	-2.26	1.77	See Figure 6
Sample size	435	2,937 (18,810)	12	320,816	See Figure 7 & Figure 8
Sample age	436	3.45	1	6	See Figure 9
Sample ability	437	2.284 (.64)	1; low ability sample	4; highly selective sample	See Figure 10
Number of questions	298	41 (31.75)	3	240	See Figure 11
Maximum allowed minutes	179	46 (40.51)	1.5	195	See Figure 12

Figure 7 Number of studies by amount of participants, only including those with less than 1,000 participants.

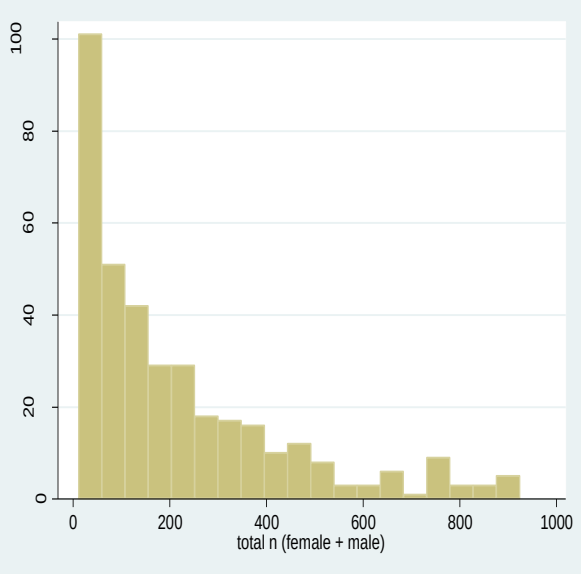


Figure 8: Number of studies by amount of participants, only including those with more than 1,000 participants.

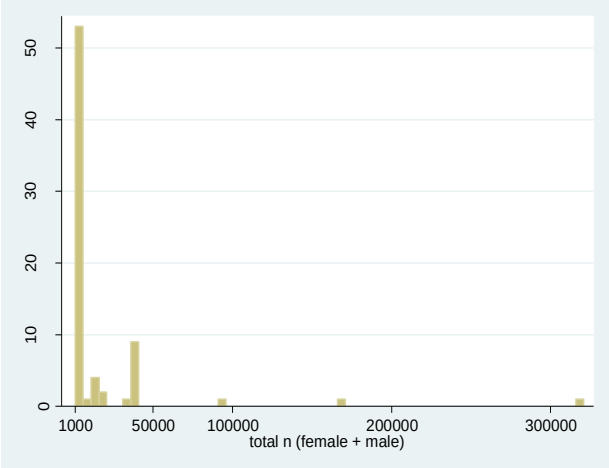


Figure 9 Number of studies with participants based on the age group of the participants. 1 signifies pre-school, 2 elementary school, 3 middle school, 4 high school, 5 college and 6 the general population.

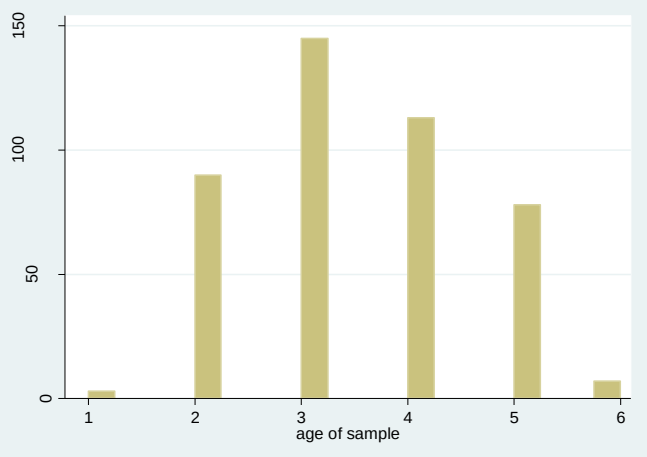
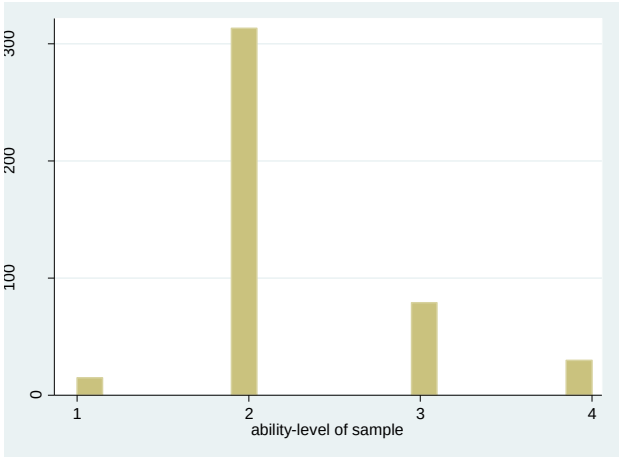


Figure 10: Number of studies with participants based on the ability-levels. 1 signifies low ability, 2 general ability, 3 moderately selective and 4 a highly selective sample.



There is a wide range for both test length variables of minute and question amount. The median number of questions is 34 and the median maximum amount of allowed minutes is 35 – see Figure 11 & Figure 12. Both the amount of questions in a test and the set duration in minutes are proxies for the test length. From Figure 13 it is observable that there is, as would be expected, a positive relationship between the two measures of length: a test with more questions generally takes longer[5].

[5] The observation that can be seen in the lower right corner of the figure refers to a test where an 'as many as you can' time format was used. Participants had 8 minutes to answer as many of the 240 questions as they could. This refers to article no. 117 from LeFevre, Kulak and Heymans (1992). It is considered with the other 'as many as you can' studies in the data analysis.

Figure 11: Histogram number of questions dataset

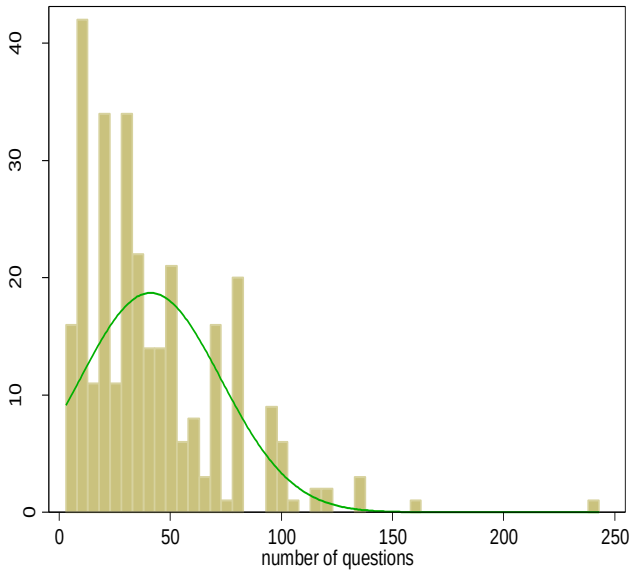


Figure 12: Histogram maximum minutes allowed dataset

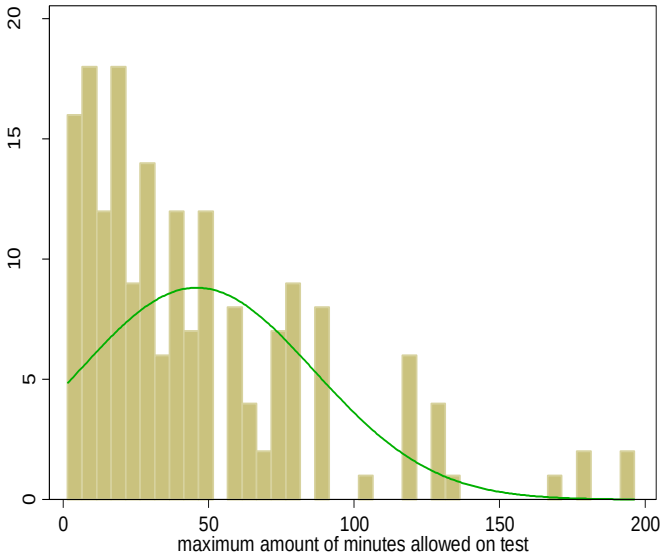
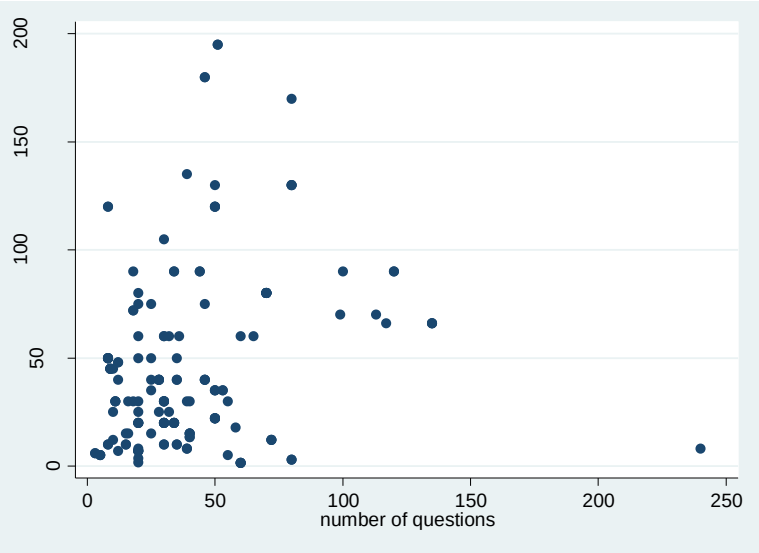


Figure 13: relationship number of questions and maximum amount of minutes



Insights in other characteristics of the tests and samples used in the studies under analysis can be found in Appendix 8.2. I would like to highlight the measure for the 'depth of knowledge' variable. This is a 1 to 4 scale variable and observations were available for 120. However, the highest level of difficulty was not tested in any of these studies and only 7% included items with level 3 (the second highest). Since literature indicates that there is a positive correlation between depth of knowledge and gender performance this may limit the external validity of our findings (see section 2.1.3). The economically relevant and most gender diffused more advanced math levels seem underrepresented.

Furthermore, the majority of studies is based on U.S. data. Aware that country differences may exist, this could also limit the external validity of our findings (see subsection 2.1.1). The difference between world regions is looked into further in section 5.2.4. These findings indicate that the impact of test length on the math gender gap differs across world regions.

4 Empirical strategy/methodology

To assess the hypothesis that males and females respond differently to test length increase, multivariate regression analysis is used. The model is of the following form:

$$d_i = \alpha + \beta_1 x_{1i} + \beta_j X_{ij} + \varepsilon_i$$

Since we only have the data available from the studies in the dataset, rather than the population, $\hat{d}_i$ is the *estimated* dependent variable of the analysis. $\hat{d}_i$ represents the outcome of interest: the standardized raw effect size of a study. Furthermore, every observation i corresponds to an observed study, rather than scores from individual participants. A given sample size in the analysis therefore refers to the amount of studies, rather than the amount of actual participants. Additionally, the estimated coefficient of test length is b_1, x_{1i} represents the number of questions, respectively the test length in minutes of study i. The relation between various independent variable values, vector X_{ji} are denoted by the vector b_j, where *j* signifies the different control variables**.**

When considering that $\hat{d}_i$ is an estimate of the true value d_i, and need not equal it, one ought to be aware of a potential for error in the regression that is not accounted for by ε_i. The difference between the observed and fitted value of the eventual regressions is captured in the error term ε_i , but an additional part of the measurement error comes from this aspect, i.e. the actual error is $\varepsilon_i + u_i$. If the gender gap is structurally overestimated (so $u_i > 0$) or underestimated, there is bias in the estimated gender gap. However as I am interested in the impact of test length on the gender gap, in other words the slope of the fitted line, given u_i is even across observations' gender gaps and test lengths, this will still yield a consistent estimate of the effect of test length on a gender gap.

The two test length measures that are considered throughout the analysis are the core independent variables and considered as two separate regressions, with b_2 to b_j added progressively as the model expands in the regression estimate:

$$(1)\ \widehat{Gender\ gap}_i = Constant + b_1 Amount\ of\ questions_i + b_j X_{ij} + \varepsilon_i$$

$$(2)\widehat{Gender\ gap}_i = Constant + b_1 Duration\ in\ minutes_i + b_j X_{ij} + \varepsilon_i$$

Throughout the data analysis, additional variables are added to determine whether the estimated coefficient of test length b_1 is sensitive to control variables. Several characteristics of each sample and study were already coded in the original dataset as variables that may influence a math gender gap. These additional variables include amongst others the publication year, test type, the world region of participants and the format of the test.

Missing variables and studies with potential problems are accounted for as the model's estimate is checked for robustness in section 5.1. Also inverse variance weights[6] are added to account for the presumed higher quality estimates of larger studies. The second part of the data analysis considers further sensitivity analyses to assess whether the strength and direction of the test length impact on a math gender gap alters in certain cases.

The remainder of section 4 is dedicated to assessing internal validity of the data (4.1), a method for dealing with missing data (4.2 & 4.3) and the comparability of shorter and longer tests (4.4).

Box 2: Introducing inverse variance weights.

Inverse variance weights are used to emphasize the difference across observations, rather than weighting them all equally. Observations' weights are calculated as follows:

$$weight_i = \frac{1}{standard\ error_i^{\ 2}}$$

The underlying assumptions is that the smaller the standard error, the more precise the effect size. The standard error tends to decrease as more subjects within a study are available.

In effect of using these weights, studies with larger sample sizes are preferred.

4.1 Assessment of internal validity of the test design

A large randomized trial would be the golden standard to test the hypothesis of test length bias in favour of females. However, since we use observational data on a large set of studies, the data used is only as good as these secondary sources. My major concern therefore stems from this use of secondary data, as I do not have direct insight in the quality of the studies used as observations. Any potential selection bias, limitations of reliability and validity, or measurement errors that might exist within this studies - I am unaware of and hence could not account for. Larger studies, which are generally considered to be superior to smaller, are given extra weight in the regression. Noteworthy, the negative correlation of the math gender gap on test length is significant in the weighted version when the amount of questions is our length variable. Weighting, however, need not capture all quality differences between studies, as was evident from the afore mentioned navy article – elaborated on in subsection 3.2. I consider this aspect of being based on secondary data the largest limitation to the validity of this thesis' design. Before analysing the data with OLS in section 5, I will consider the internal validity of our data by discussing the relevant assumptions underlying multivariate regression.

4.1.1 Normality assumption

The assumption of conditional normality of the dependent variable is tested via the estimated residuals of the error terms. A visual representation is available in Figure 14 and Figure 15. Though they are spiked, we see a roughly normal distribution, and given the considerable sample size I do not consider this a problem.

[6] See Box 2 for an elaboration in inverse variance weights.

Figure 14: Residuals from the simple linear regression of the number of questions on the standardized gender difference

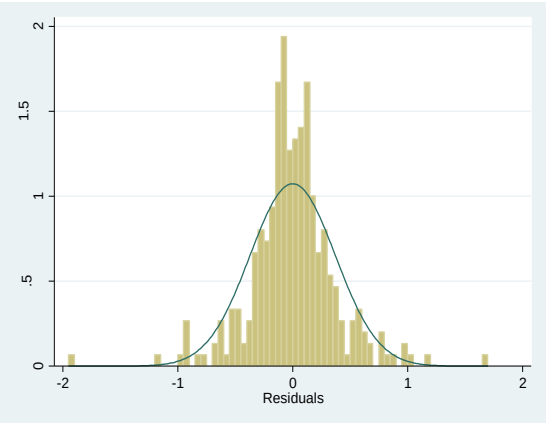


Figure 15: Residuals from the simple linear regression of the number of questions on the standardized gender difference

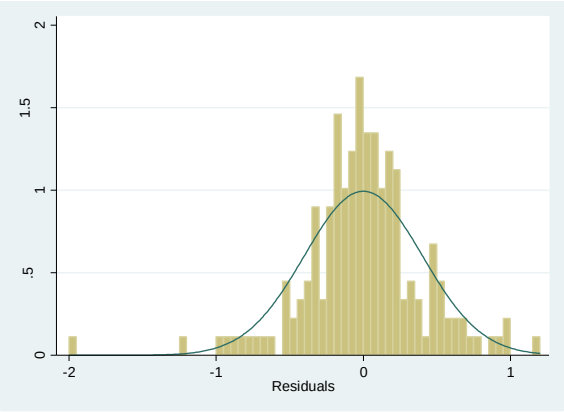


Figure 16 and Figure 17 plot the estimated residuals of the error terms of the more extensive model (corresponding to column (7) of Table 5). In these cases the distribution of the error terms are particularly peaked. Again, given the considerable sample size I do not consider this a problem.

Figure 16: residuals from the full model regression of the maximum amount of minutes of tests on their standardized math gender gap (with all controls, weighted observations, excluding article 87 and clustered standard errors).

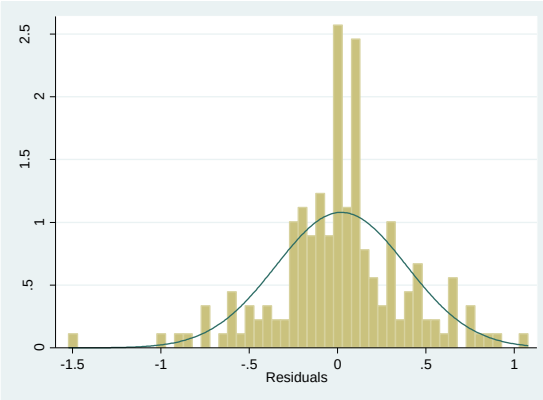
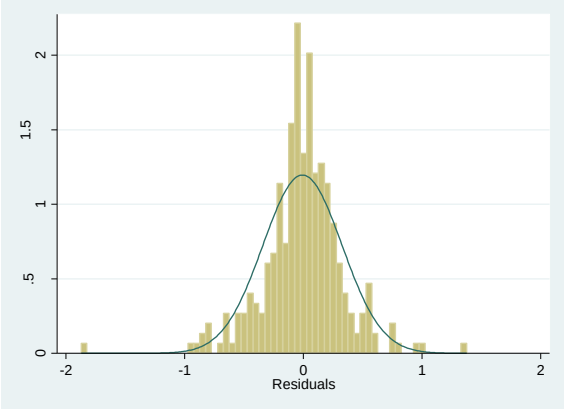


Figure 17: residuals from the full model regression of the question amount of tests on their standardized math gender gap (with all controls, weighted observations, excluding article 87 and clustered standard errors).



4.1.2 Homoscedasticity of the error terms

I used White's test to check for homoscedasticity of the simple linear regression of the two test length independent variables on the standardized gender difference. Based on this test I could reject the hypothesis that there is heteroskedasticity in both simple linear regressions. However, for the extended model including all controls (unweighted observations, standard errors not clustered) I had to reject the null-hypothesis of homoscedasticity. The variances are no longer the same and constant across the regression. This compromises the ability to draw inference from OLS and introduces efficiency issues.

4.1.3 Linear relationship between the independent and dependent variable

A problem with using linear regression without transforming variables is that the regression analysis may under-estimate the true relationship. When generating the squared values of the 'number of questions'-variable, a fitted regression of the standard difference seems more linear at a glance – though also containing more variation. This can be seen from comparing Figure 18 to Figure 19: when adding a cubic term of our length measure, in the number of questions case the fitted line becomes steeper It becomes flatter for the maximum amount of minutes measure, observable from Figure 20 to Figure 21. As the data is scattered it is hard to check for this assumption by only using visuals, in the data analysis of subsection 5.2.5 the impact of adding polynomials is therefore also considered in the regressions.

Figure 18: Scatterplot and fitted regression on standardized difference on squared number of questions

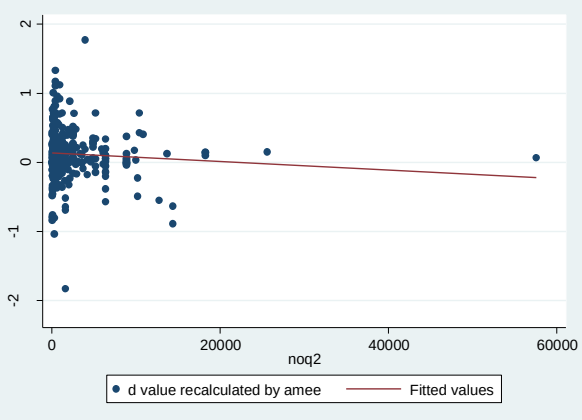


Figure 19:: Scatterplot and fitted regression standardized difference on number of questions

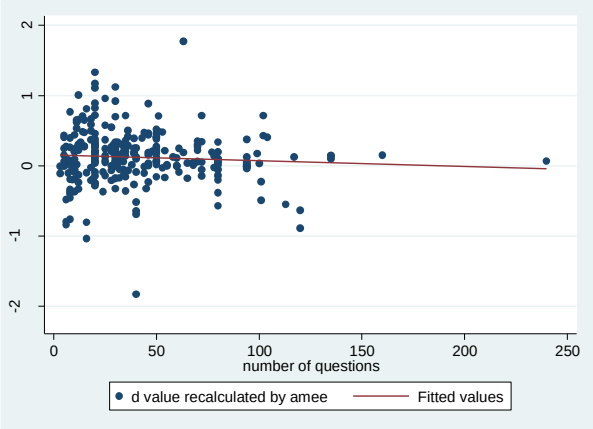


Figure 20: Scatterplot and fitted regression on standardized difference on squared maximum amount of minutes

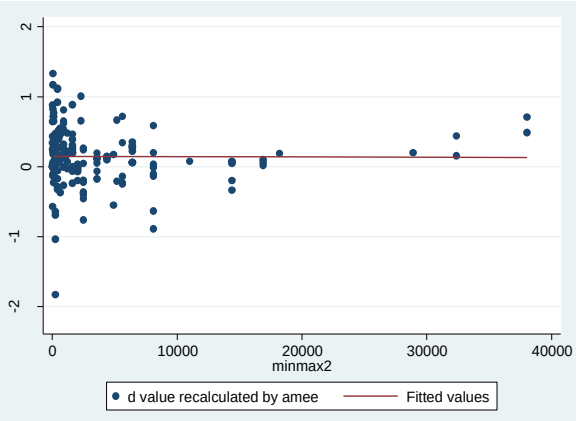
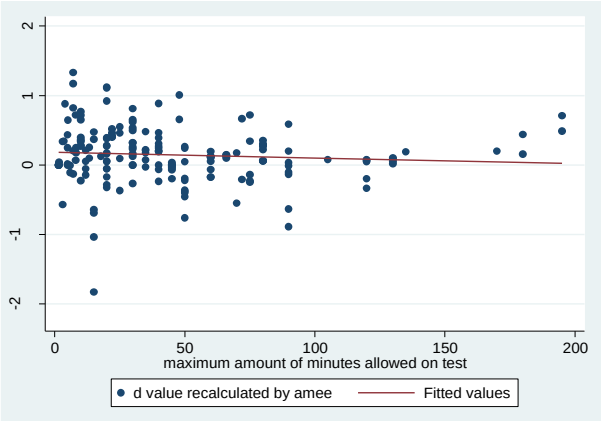


Figure 21: Scatterplot and fitted regression on standardized difference on maximum amount of minutes



4.2 Availability of information on test length

Not all observations' test lengths could be determined. When using a t-test we have to reject the hypothesis that the mean gender gap for the group of observations where these variables are known is equal to those where the test lengths are unknown at a 10% significance level. The mean standardized difference for the 143 observations of which we do not know the number of questions is 0.057, as compared to 0.12 for the 298 where it is known. The mean value for the 262 observations of

which we do not know the test length in minutes is 0.07, the mean value of the 179 for which we do have this value is 0.15.

This is potentially problematic, may indicate selection bias and particularly limits the external validity of the findings. Selection bias implies the sample that was used to obtain the final dataset is not randomly selected from the original data set (that I assume here to be a random obtainment of the studies on the math gender gap). This could be the case when certain observations, namely those where boys scored better, were easier to come by than others. For the 149 observations for which I know whether they were made under a time limit or not – I do not need to reject the null hypothesis of equal means as compared to the group of missing variables for the math gender gap.

The different means for the test length variables are illustrated in boxplots in Figure 22 and Figure 23, where the observations with known values (group 1) seem more spread out than those for which data was missing (group 0). The significant mean difference visually seems quite limited and merely caused by a higher spread in group 1 To enquire what may underlie the different outcomes of the two groups, I consider whether there is a confounding variable. there may be overrepresentation of a subset of studies with a variable that both determined boys obtaining higher scores and implied an ease of access to test length. Possibly for tests with certain characteristics it was harder to obtain test length data. Such factors could also be of relevance to the standardized difference and could indicate a cause for the unequal means.

Figure 22: Boxplot of two groups of observations: those with data available for the number of questions (group 1) and those with missing data (group 0)

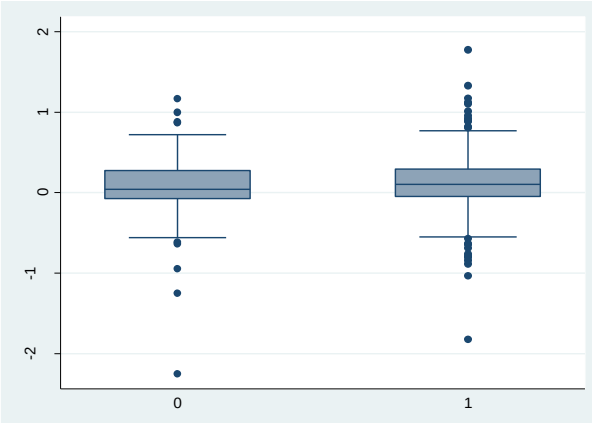
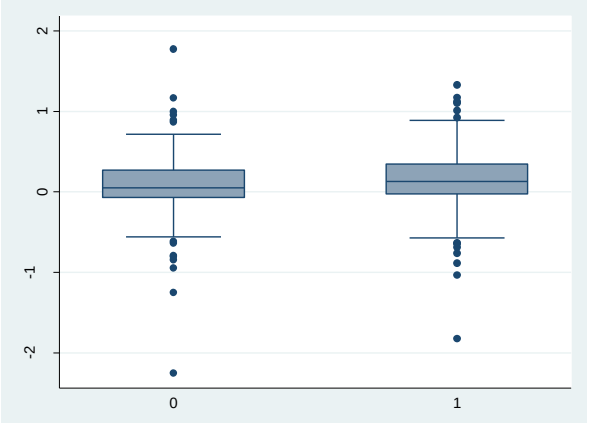


Figure 23: Boxplot of two groups of observations: those with data available for the maximum amounts of minutes of the test (group 1) and those with missing data (group 0)



I first considered the world region aspect. Indeed, there are significantly more U.S. studies in the sample that has test length observations. However, this variable does not significantly impact the gender gap so cannot be the cause. Second, and surprising, is that there is a significantly larger amount of low stakes tests in the final dataset as compared to the unobserved studies. This is surprising as literature indicates that low-stake tests should favour girls, see section 2.4.2, so the higher d-value that is obtained for the observed group should not be retraceable to this variable. I thus considered chi-squares for the four variables that significantly impact the standardized gender difference according to Lindberg et al.'s (2010) original article. Chi-squares compare the expected values from a group to the actual values of another group that is assumed to be similar. In this case, I consider if the group of studies where no observation of test length is present has significantly different shares of categorical variables than the group where these test length variables are available.

Ability

For the number of questions, at the 5% level there is a significant difference between the observed and unobserved group. For the maximum amount of minutes variable there is a significant difference at the 10% level. There are relatively many highly selective studies for which data was available and slightly more moderately selective samples. Given this, the higher d-value obtained in the non-missing variable group may be driven by this.

Ethnicity

The variable 'ethnicity' implies that a sample primarily consisted of subjects from minority groups. For neither of the test length measures, the groups with versus without observations on test length have a significantly different share of ethnicity in them.

Age

Neither of the differences between observed and unobserved testlength group imply different age groups

Test type

When considering the number of questions, no test type prevailed or was only limitedly present in either group. The maximum amount of minutes measure did provide two significantly different means. Type1 ('includes multiple choice questions') was at a 10% significance level more prevailing in the group of observed data. Type2 ('includes short answer types') was different at a 1% significance level. There were more of such studies in the final dataset with observed test length, as compared to the unobserved test length sample group. Given this high significance, and the impact of test type on the math gender gap, this may be one factor driving the difference in means of the unobserved versus observed test length data.

When we alter the sample under consideration to only include observations of which it is known they had a time limit, there is no longer a significant difference in mean value between groups with and without observations for either test length measure. Here, the group with an observed number of questions (N=203) has a mean standardized gender difference of 0.13, the group with no observed question-amount (N=42) has a mean d value of 0.06. The two-sample ttest is not significant (p=0.27). When we consider the observations of timed studies, the group with missing observations (n=70) has a d-value of 0.07, the group with observations (N=175) have a mean of 0.14 – also here however the difference is insignificant. This is a positive outcome for the validity as the subset of studies under timed circumstances could imply more significant boundaries of test length. Considering these insignificant results for the timed observations with the limited distributional differences from the boxplot, the situation does not seem too problematic. However, problems do potentially exist. Internal validity may be limited as observations from our original (assumed random) dataset are not represented accordingly in the final dataset used for the analysis of this thesis. Moreover, there are limitations to generalizing the findings.

4.3 Dealing with missing data

To research whether the missing observations of the control variables were randomly missing, I used two sample t-tests to compare the mean d-values of those with and without observations. I moreover tested for equal mean values of the number of questions and amount of minutes between these two groups. The outcomes of these tests are provided in Table 4.

Table 4: Outcomes two-sample t-tests to test the null-hypothesis of equal means between groups without (marked as '0') and groups with (marked as '1') observations on the different control variables. The null-hypothesis was rejected when the p-value for the hypothesis was below 0.10.

Outcome variable:	**Group?**	**Depth of Knowledge**	**Test format**	**Test type**	**Content type**	**Stakes**	**Ability**	**Age**
Standardized difference	0	0.11 (N= 316)	0.08 (N=207)	0.10 (N=248)	0.09 (N=232)	0.09 (N=249)	-0.02 (N=4)	-.0003 (N=5)
	1	0.09 (N=119)	0.13 (N=228)	0.10 (N=187)	0.12 (N=203)	0.12 (N=186)	0.10 (N=431)	0.10 (N=430)
Reject H0?		No	No	No	No	No	No	No
Number of questions	0	44.10 (N=193)	45.45 (N=106)	46.09 (N=138)	45.98 (N=123)	42.40 (N=149)	66.5 (N=4)	66.5 (N=4)
	1	35.53 (N=105)	38.69 (N=192)	36.76 (N=160)	37.64 (N=175)	39.77 (N=149)	40.74 (N=294)	40.73 (N=294)
Reject H0?		Yes. 5%	Yes. 10%	Yes. 5%	Yes. 5%	No	No	No
Maximum amount of minutes	0	57.13 (N=117)	53.96 (N=55)	49.27 (N=77)	57.85 (N=90)	38.99 (N=64)	39 (N=4)	39 (N=4)
	1	24.64 (N=62)	42.29 (N=124)	43.32 (N=102)	33.77 (N=89)	49.71 (N=115)	46.03 (N=175)	46.03 (N=175)
Reject H0?		Yes. 1%	Yes. 10%	No	Yes. 1%	Yes. 10%	No	No

Overall we can see that tests are shorter for observations of which data was available, as compared to the missing data groups. This seems the case at a 5%, respectively 1%, significance level for the depth of knowledge and content type for both length measures. Test measures also differ at the 10% significance level for the groups of observations with and without test format availability. The same goes for one measure in the case of test type and stakes. Ability and age variables, whether missing or not, are of equal lengths. These results could bias our findings as it gives rise to questioning the distribution of our test lengths across these controls. Also weights were not available for all observations. Those observations were dismissed as weights were added to the regression in section 5.1. The impact of estimating these weights is also provided in that section..

Some of the sample sizes are underlined in Table 4, these are the amounts that were eventually coded as a dummy in the data analysis. Dummies were added if an observation of a categorical variable was missing, in order to keep the amount of observations throughout the expanding regressions constant. Since all missing control variables were actually categorical, no steps had to be taken transform continuous variables. Lindberg et al. (2010) did note, "the level of missing data for moderator variables (due to vague descriptions of mathematics measures and study procedures) made it untenable to conduct a simultaneous analysis of all moderators)" (p. 1127). Given the amount of missing data, this decision may not be fully warranted. As a visual example of this Appendix 8.3 presents a graph is to indicate how the studies are distributed differently in the two groups for the observations with data on depth of knowledge.

4.4 Comparability of short and long tests

To consider whether short and long term tests only differ in length, or also in other characteristics, I consider if short versus long tests have equal means across other variables. The same variables as in section 4.2. and 4.3. are considered. If these differ significantly, this limits the internal validity of the regression: the difference in estimated math gender gap as test length changes may stem from changes in these other variables rather than the change in length. I split the test lengths based on both the median and the mean. This seems relevant given the clustering of studies at the shorter lengths, implying a mean could be far to the left of the distribution and in the middle of short tests rather than being a legitimate split between short and long.

As expected, the medians are 8 minutes, respectively 10 questions, lower as compared to the means. For the number of questions length variable: the split of groups by means (at 41.08 questions) yields 184 short tests and 114 long tests. The split of groups by medians is at 33.5 questions and, by definition, splits the group in half (149 in both). For the minute amount, the mean is 45.88 minutes (yielding 112 short studies, and 67 long) and 35 minutes for the median. As there were six observations of 35 minute length, there are 93 observations considered 'short tests' and 86 'long tests', since those studies equalling the mean or median amount were included in the short group. Results do not differ when including these six studies to the large test group. Again chi-squares will be used for these measures since both are categorical variables (short/long tests and e.g. inclusion/exclusion of test type). The outcomes are provided per variable.

Depth of knowledge

The question amount measure split by means indicates there are significantly more often items with the lowest depth of knowledge level included in longer tests. Which implies these are relatively easier. However, the minute amount measure (for both splits) indicates the opposites: there is relatively more inclusion of level 2 depth of knowledge items in long as in short tests and fewer level 1 (the lowest) level in longer tests. These mixed results may indicate there is some clustering of level 1 and level 2 included items at the border of short and long, depending on which measure is used. Other test levels and the number of questions median split do not indicate significantly different amounts.

Test format

This variable considers the format (computerized, behavioural/oral and paper and pencil) of a test. When we consider test length indicated by question amount (mean split), we find there are significantly less oral/behavioural format tests in longer tests (at a 5% level) as compared to short tests. Other formats are similar. The minute amount measure indicates (for both splits) this as well: there were none oral/behavioural longer tests, though a 10% share of short tests was of this format (at 5% significance level for both splits). There are significantly more paper & pencil format tests (at 10% for mean split, respectively 5% level) for those tests that are longer. The other formats were not significantly more or less common.

Test type

This variable considers the inclusion of answer types in tests. For the mean question split it is found there are relatively many multiple choice and short answer items included in longer test (both at 5% level). The median split question measure indicates the same significant findings for both answer types (at 1% and 10% respectively). There is no significant difference in inclusion of open response items. For the other length measure, in both splits, there are significantly more multiple choice and open

response items included in longer tests. No significant differences for short answer items are found. These findings imply that tests with more questions are more likely to include all types of answer items is not really surprising: if you have more questions available in total, you may be more inclined to vary with answer type.

Content type

The different length measures and splits yield mixed results for more or less prevalence of content items in different groups of length. The mean split of questions indicate there are relatively little algebra items included in longer tests, though relatively many measurement items (both at 10% significance level). The minute of questions split by means indicates number and operations items were included little in longer test (at 1% level), algebra items (contrary to the earlier finding) relatively often (1% significance level) as compared to shorter tests. Data analysis & probability items were also prevailing more in the longer tests. The median split for length in minutes grouping indicates the same findings as the mean split of this test length. Given the significant findings in most of the content types, this is of potential concern, though the mixed results imply little indication for the direction of bias.

Stakes

The splits for the question amount measure did not imply significantly more or less low or high stake tests in either of the length groups. When comparing short and long tests based on their test length in minutes, we do find there are significantly more high-stakes tests in the long group as compared to the short groups, where there are relatively more low-stakes tests. This is a significant difference at a 5% level.

Ability

The mean measure of the number of questions and both splits for the minute maximum imply that there are significantly fewer moderately selective and highly selective samples (but quite many samples of general ability) in the longer tests as compared to shorter tests (at 1% significance).

Age

The overall chi-squared estimate in all measures but the median split by question amount imply a different distribution. The numbers indicate that short tests are primarily made in college age groups (and in some measures elementary school), whereas the longer tests are more prevailing for middle and high school subjects. Significance levels range from 1% to 5%

Nationality

The different test length measures imply different findings. The question amount where groups are based on the mean split indicate there are relatively many Canadian samples in the longer tests, and relatively more from Europe, Oceania and Asia in the shorter studies (at 5% significance level). The mean minute findings indicate (for the mean split) Europe, Oceania and Asia were overrepresented in short tests. The median split indicates long tests in the data set are more often from Canada, Oceania and the Middle East than would be expected. The U.S. and Europe take up a larger share of the short test sample.

Ethnicity

The median split for the question amount length implies there are relatively more primarily euro-American samples (rather than primarily minority subjects) in the shorter tests as compared to in the longer tests (at 10% level). The maximum minute amount also indicates there are relatively many

primarily minority groups represented in the longer tests and relatively many primarily euro-American samples that made short tests, and relatively more primarily minority samples making longer tests.

The oftentimes significantly different characteristics of these groups is potentially problematic for the internal validity of our results. When noting that age, ability, ethnicity and test type proved of significant impact in the Lindberg et al. (2010) meta-analysis, these significant different prevalence are of particular concern. Their direct impact on the standardized difference combined with their correlation with test length may make them a confounding variable and blur the relation of impact of test length on impacting the gender gap. Given that samples of primarily ethnic minorities implied a lower d-value according to Lindberg et al. (2010), and that there are more primarily minority samples in longer tests, this is an indicator of potential overestimation of our results. Given the d-value is lower for the general population (see Lindberg et al., 2010) and these are overrepresented in longer tests, this is also of potential concern. Age groups in college and elementary school are overrepresented in short tests (both corresponding with higher math gender gaps, of respectively +0.06 and +0.18) and longer tests have higher shares of in middle (+0.00) and high school (+0.23). This yields inconclusive findings. Also the mixed results of the test type differences of long and short tests do not imply strong concerns. All in all, we should be cautious in interpreting the results of this thesis. Given that these factors are included as control variables in the fuller regressions and considered interacting with test length, it is largely accounted for in the analyses.

5 Estimated test length effects

In this data analysis, a total of 435 studies are used as observations. These studies combine the test results of 1.277.598 subjects – less than the original article by Lindberg et al. (2010) due to the elimination of six studies as explained in subsection 3.3. The current section provides statistical insights of the impact of test length on math score differences between males and females.

Two proxy-variables are considered for assessment of the test length: the length in minutes and the number of questions. These provide insights into the length and are therefore considered in greater detail than the binary measure of a test having a time limit or not. Lindberg et al. (2010) considered this presence of a time limit and found a non-significant coefficient for the moderator variable of a test being 'timed'. Also in this thesis' final dataset, no significant impact on the gender gap is found when a test is timed versus when it is not. The relationship between the length measures and the 'timed' variable is considered in more detail in subsection 5.2.2, given the potential impact of setting a time limit on a test with any given length. Figure 24 and Figure 25 provide a graphical representation of the relationship between the maximum amount of minutes allowed on a test, respectively the number of questions, and the standardized difference between male and female math test performance. It can be observed both present slightly negative lines and have scattered data.

Figure 24: Fitted scatterplot maximum amount of minutes allowed on standardized math gender difference

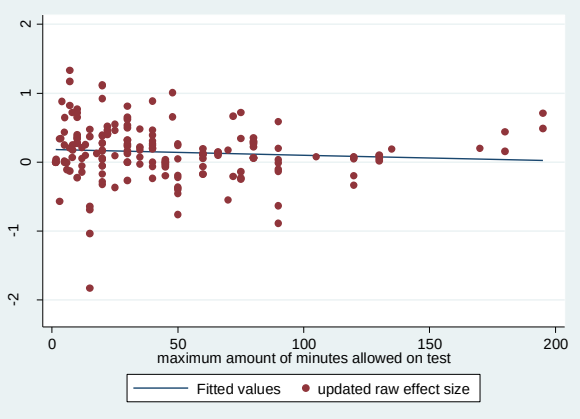
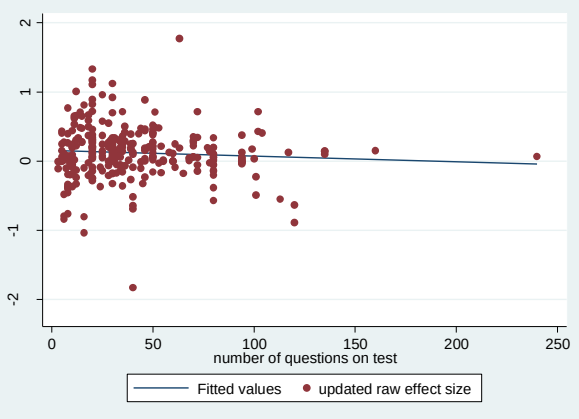


Figure 25: Fitted scatterplot number of questions on standardized math gender difference



The regressions' estimates will be provided in detail in next subsection. All estimates of test length impact on the standardized gender gap, except two, are not significantly different from zero, though all indicate small negative values. This could be a weak indication of a small female advantage as test length increases. The simple linear regressions indicate a non-significant average *decrease* of 0.0008 for the standardized math gender gap per added minute or question. E.g. when a test lasts 20 extra minutes, the standardized difference decreases with 0.016. Adding all controls on test and subject characteristics and accounting for the weights available yields an influence per added minute or question on the gender gap of -.002. This is significant at a 5% level for the questionamount measure.

The remainder of this section is structured as follows: the overall regression estimates are provided (5.1), Via further sensitivity analysis (5.2) I first consider the potential moderator variables as were of significant impact on the math gender gap in the Lindberg et al. (2010) paper (5.2.1). Then, I consider including only studies from which it is known they were timed (5.2.2), I account for the missing variables (5.2.3), world regions (5.2.4) and polynomials (5.2.5) thereafter.

5.1 Regression estimates

The estimates from ordinary least squares regression are presented in Table 5. As one article sometimes generated multiple effect size estimates since they included several studies, I cluster standard errors on the article level. These studies frequently have characteristics in common: the same set of children, the same test used, the same school district. Not accounting for these within-group similarities would underestimate the variance of the estimate. When clustering standard errors, independence across articles is still assumed but correlation within articles is accounted for.

The first column of Table 5 describes a simple model - the only explanatory variable for the math gender gap is the test length measure. Where a non-significant parameter of -.0008 is found for both length measures of interest. In order to consider the real effect of test length on the math gender gap (our variable of interest) control variables are added gradually. The dataset of Lindberg et al. (2010) allows for exploitation of the many variations of the studies made explicit in the dataset. The additional variables are added per group rather than individually to the simple regression – the control variables are clustered for being either student background or test features. In Appendix 8.6 an expansion of the regression per individual variable can be found.

I decided to first add whether a test was timed (column 2), as this is directly connected to test length via actual perceived time pressure. Due to this importance, this control variable is added individually rather than as part of a larger group. Controlling for this does not alter the estimates. Thereafter, a time trend (column 3) was added, followed by adding the controls in two steps: the student background characteristics (nationality, age and ability level, column 4) and then the test characteristics (the stakes, characteristics, format, and content type, column 5). Although adding these three sets of controls alters the size of the coefficients in the range of -.0002 to -.0010, they remain insignificant and negative for both test length measures of minutes and questions. Also individual addition of the variables does not yield a significant parameter for our variable of interest (see Appendix 8.6) – weights are not considered in the variable-by-variable approach.

Column (6) accounts for the weighing of variables. As weights were not available for all observations these were in general dismissed when moving to regression (6) and (7). As a weighting method inverse variance weights are used. In the inverse variance method the weight given to each study is the inverse of the variance of the effect estimate (i.e. one over the square of its standard error)[7]. Thus larger studies are given more weight than smaller studies, as these tend to have larger standard errors. Column (7) excludes the heavy-weight, badly designed for this thesis' aim, study of the navy article by Held, Alderton, Foley, & Segall (1993) elaborated on earlier in section 3.2. The weighted outcomes excluding this study yielded -.0019 (p=0.15) where the maximum allowed amount of minutes was used as an estimator for the standardized difference. For the number of questions as the variable of interest a parameter value of -0.0022 was estimated (p=0.03). I think regression (7) thereby becomes the most convincing regression and hence estimate of the thesis. The large rise in R-squared from (5) to (7) – though not perfectly comparable as thirteen observations are excluded moving from (5) to (7) plus the exclusion of the navy article – also indicates an increase in the predictive power of this regression as a whole.

[7] Please see box 2 at the start of section 4 for an elaboration on inverse variance weights.

As weights were not available for all observations these were dismissed when moving to regression (6) and (7). However, those missing were observations for which a d-value and total sample size were known. The standard deviations and sample sizes per gender were missing. When I did replace these missing weights by assigning them the average of the weights three studies above and three studies below them (sample size wise), the regression estimates for test length alters. For the number of questions measure .0020 (standard error .0000) (n=129) is the new coefficient. This is highly significant and disturbing given the positive direction. For the maximum amount of minutes, the estimate also becomes significant, though the size of the parameter does not alter much as compared to excluding the unweighted studies: -.0018 (standard error .0001) (n= 178). This strengthens the finding that for the regression of the maximum amount of minutes as a length proxy, a significant impact on the standardized math gender gap of around -.002 per added minute can be justified. The alteration in direction for the number of question measure seems to come from six studies (all from the same article), all with top 10 size samples. The impact of these observations puts pressure on the claim that regression (7) is the best estimate for the impact on gender difference of the number of questions – as it is not robust to the adding of observations with missing weights. Nevertheless, given these weights are only estimates on the basis of their sample size, rather than having anything to do with their standard errors, this may not be too severe.

Column (8) is added to exclude those studies where potential problems occur. This to determine whether excluding the observations constructed worse indicate a different effect of test length than we previously found. Some studies suffered from the following problems:

1. Schoolyear data or generally odd grades. So scores determined based on year-round grades, hence no test length could be determined. I considered grades 'odd', when the length was irretraceable by the study's nature: they were based on multiple tests, or 'sorting numbers' rather than math, variables were standardized rather than using original test scores;
2. There was a mismatch between the numbers the original meta-analysis data-set and the original study;
3. The time test length was an estimate of the author;
4. Data of the original study was left out - this would happen for good reason in the original meta-analysis based solely on the gender gap but implies a selection that may not be justified given the current hypothesis;
5. The test length indicated includes more than just the time to make the test – it for instance also includes time to fill in another questionnaire. This makes it an imprecise measure of time;
6. All studies based on behavioral observations.

Excluding the observations suffering from these potential problems provide insignificant coefficients of respectively -.0010 and -.0007 for the maximum minutes of a test and the number of questions measures on the math gender gap. I am a bit unsure whether at the time of keeping track of the issues that may prevail in studies I was fully consistent and aware what would be problematic. Hence, although I think regression model (8) can be of some added value, I do not want to put too much weight on these findings.

Furthermore, as was discussed in section 4.4, a split can be made between short and long studies based on their means and medians. Doing so creates a binary variable, where a test is considered either short or long. Taking the minute amount as the length measure of interest, the d-value for long tests as

compared to short differs significantly at the 5% level. Short tests indicate a d-value of 0.20, long tests one of only 0.05. When splitting observations into short or long tests based on the median test length in minutes, this effect is similar: the estimated d-value for short tests is 0.21, that of long tests is 0.07 (also at a 5% significance level). These findings hold when including only timed studies, but are insignificant when the question amount is used as the length measure. Given that I also estimated significantly different characteristics between these short and long groups, I would be cautious to trust these outcomes fully. A final general finding is that when only considering those tests where there was an 'as many as you can'-timing element: e.g. 'answer as many out of 40 questions as you can in 10 minutes', I do not find a significant impact.

What strengthens a conclusion for a potential impact of test length, comes from only including studies with a total amount of participants of 30 or above. This yields a coefficient estimate for the amount of questions on standardized test length of -.002 at a 5% significance level. The same value, though not significant, is found for test length in minutes as an independent variable. When considering that greater sample sizes are given more weight as they are considered more valid experiments and this yields significant results (as we know from regression (6) and (7) of gender gap on the number of questions), this is not very surprising.

In summary, the results indicate that for all regression specifications in Table 5 the direction of the estimated coefficient for test length (be it in minutes or questions) remains negative – which implies robustness. There is generally a small negative impact (in the range between -.0002 and .0022) on the raw effect size when incorporating other factors that may impact the standardized difference between boys' and girls' results. Particularly the significant implication of every additional question implying a significant -.002 impact on the math gender gap in the weighted elaborate model is an interesting find. Given the lack of significance when excluding potentially problematic observations from the regression as well as the positive direction when estimating missing – results are most of all mixed. Whether a female favoured bias occurs as test length increases is unsure. In the upcoming subsection I look into some specific groups and factors to estimate if effects may be less or more prevailing under different circumstances.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Effect on standardized gender gap								
Maximum minutes allowed	-.0008 (.0010)	-.0007 (.0010)	-.0010 (.0011)	-.0002 (.0012)	-.0005 (.0013)	-.0016 (.0013)	-.0019 (.0013)	-.0010 (.0007)
Number of observations	179	179	179	179	179	166	165	117
Adjusted R-squared	.0012	-.0022	.0191	.2029	.2407	.7465	.6313	.8366
Time limit	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year trend	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Student background controls	No	No	No	Yes	Yes	Yes	Yes	Yes
Test characteristic controls	No	No	No	No	Yes	Yes	Yes	Yes
Weigh observations	No	No	No	No	No	Yes	Yes	Yes
Excluded observations	No	No	No	No	No	No	No	Yes
Effect on standardized gender gap								
Number of questions	-.0008 (.0009)	-.0008 (.0009)	-.0010 (.0008)	-.0002 (.0007)	-.0004 (.0008)	-.0021** (.0010)	-.0022** (.0010)	-.0007 (.0007)
Number of observations	298	298	298	298	298	285	284	217
Adjusted r-squared	.0014	-.0029	.0244	.2305	.2400	.7553	.5319	.6885
Time limit	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year trend	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Student background controls	No	No	No	Yes	Yes	Yes	Yes	Yes
Test characteristic controls	No	No	No	No	Yes	Yes	Yes	Yes
Weigh observations	No	No	No	No	No	Yes	Yes	Yes
Excluded observations	No	No	No	No	No	No	No	Yes

*Table 5: Regression estimates of the standardized math gender gap on test length measured as allowed minutes or the number of questions while gradually adding controls. All standard errors are clustered by their original article. Column (1) provides a simple linear regression. As of column (2) a dummy of a time trend is added, as of (3) a linear time trend is added, as of (4) the background characteristics of ability, age and nationality are accounted for and as of (5) test characteristics (depth of knowledge, stakes, answer type (multiple choice, short answer items and open response items) test format and content type (geometry etc.)) are included in the model – reaching a full model. For robustness analysis I then account for the inverse variance weights of observations (6). (7) uses weighted values, but excludes the observation of article 87, its reasons elaborated on in subsection 3.2. Thereafter I consider what happens when excluding observations that were behavioural observations or had noted problems in column (8). ** indicates p<0.05.*

5.2 Further sensitivity analyses

In this section there will be attention for a more detailed view on when and to what extend test length has an impact on the differences in gender performance in math. At first (subsection 5.2.1.) I will consider those factors that significantly influenced the math gender gap in the article by Lindberg et al. (2010). These are considered as moderator variables and are test type, ability, ethnicity in the U.S., studies and age. I am curious whether these different levels or groups respond differently to length increases. Next considered will be the regression estimates of standardized difference on test length using only tests with a time limit (5.2.2.). Then I look into the effects of missing test-length variables (5.2.3.). Finally, I will consider whether studies with populations from different world regions respond differently to different lengths (5.2.4.) and how the estimates alter when adding polynomials (5.2.5.).

5.2.1 Accounting for potential moderator variables

5.2.1.1 Accounting for test type

Different mathematics tests are used in the different observed studies. The test type variable is set in three types: type 1 signifies that the shortest type of questions are included in the test (multiple choice), type 2 relates to the inclusion of short answer items and type 3 to the inclusion of open response items. It is reasonable that certain types of questions take more time to answer than others. This would be observable via a relationship between question type and test length. The correlation with question type and the amount of questions indeed shows this direction: Type1 (shortest): 0.20, Type2: 0.11 and Type3 (longest): -0.17. Ergo, inclusion of shorter answer formats correlate with less answers on a test, a test including longer answer types tends to consist of less questions. As we know from section 4.4, longer tests also have significantly higher levels of inclusions of both type 1 and type 2 items as compared to short tests. The reason for further analysis is that there was a relevant impact of the test types on the math gender gap estimated by Lindberg et al. (2010): “Tests with a higher proportion of multiple choice and open-ended items yielded smaller gender effect sizes, whereas tests with a higher proportion of short answer items yielded larger gender effect sizes.”

Table 6 provides an overview of the impact of test type. By adding an interaction term, it can be assessed if a particular combination of type with length adds significant meaning above the test and question type variables in explaining the standardized gender difference. For the ‘number of questions’ length measure, we see all but the open response dummy estimates are significantly different from zero. Given that the constant term in the regression is estimated (significantly different from zero) at -.188, the significantly positive test length coefficient of .008 is overruled or nearly compensated when including either interaction with type1 (-.01) or type2 (-.008) items. A test including multiple choice items as well as short answer type items needs 52 questions to eliminate the full male preferred math gender gap (-.188+.425+.275+(52*(.008*-.0104-.0075)=0. No other results of this test length measure are significant.

Table 6: all regressions only include timed studies, the regressions also include dummies for year of publication. Standard errors are in parenthesis. (1) and (4) relate to type 1: Includes multiple choice items, (2) and (5) to type 2 Include short answer items (3) and (6) to type 3: Includes open response items. All regressions are weighted and standard errors are clustered by article.

Dependent variable: standardized gender difference	**Number of questions**	**Maximum amount of minutes**
Sample size	149	91
R-squared	0.198	0.172
Test length measure	.0080***	-.0032 (.0058)
Includes multiple choice items dummy	.4248*** (.0806)	.0308 (.1974)
Includes short answer type dummy	.2752*** (.0835)	-.2344 (.1814)
Includes open response type dummy	.1100 (.1093)	-.0706 (.2114)
Test length * multiple choice	-.0104*** (.0025)	-.0010 (.0044)
Test length * short answer	-.0075*** (.0029)	.0052*** (.0019)
Test length * open response	-.0036 (.0026)	-.0008 (.0027)

5.2.1.2 Accounting for ability

Lindberg et al. (2010) found ability of the subject group to have a significant impact on the standardized gender difference. For samples of the general population, d = +0.07, but d = +0.40 for highly selective samples. There are four ability levels defined: low ability, general ability, moderately selective and highly selective. In line with their categorical character, I set them in four dummy variables and estimated a regression model including all. For the test length variable measured as minutes, significantly different values from zero are at times estimated for the coefficients of length, ability level and interaction on the standardized gender gap.

Of some concern is the positive direction of the test length coefficient here (of 0.01), which we also saw in the previous estimate on test type. Only for the moderately selective sample, we see an interaction term that is significant and negative, which could mitigate this effect. As an illustration: for a study with a moderately selective sample, a test would need to contain 226 ((0.8163-0.4630[8])/(0.0103-0.01186) questions to fully eliminate the math gender gap. A highly selective sample is estimated to start off with a d-value of 0.24 given no questions and a gradually rising gender gap as test length increases. For a general ability, it is estimated at 46 questions that the math gender gap in favour of girls switches to boys. These latter two findings undermine our hypothesis.

Table 7: regressions with interaction effects for 'ability' variable. All are weighted and standard errors are clustered by article number. Standard errors are between brackets.

Dependent variable: standardized gender difference	**Number of questions**	**Maximum amount of minutes**
Sample size	282	163
R-squared	0.3299	0.2955
Test length measure	.0005 (.005)	.0103* (.0057)
General ability dummy	-.3571 (.2692)	.1719 (.3367)
Moderately selective sample dummy	.2241 (.2488)	.8163** (.3387)
Highly selective sample dummy	.3555 (.3291)	.7075** (.3540)
Test length * general ability	.0053 (.0055)	-.0078 (.0058)
Test length * moderately selective	-.0011 (.0051)	-.01186* (.0059)
Test length * highly selective	-.0037 (.0057)	-.0099 (.0060)

[8] -0.463 is the estimated constant in this regression

5.2.1.3 Accounting for ethnicity in the U.S.

As the binary variable 'sample population being primarily of any minority group' versus 'primarily euro-American sample' in the United States significantly impacted the math gender gap – "Samples composed mainly of whites showed d = +0.13, whereas for ethnic minority samples, d = –0.05" - it is of interest to consider if test length impacts these groups differently. When including the observations in regressions with test length, the ethnicity dummy and an interaction term, we find significant impact of test length and the minority group measure, but not of the interaction term – as can be observed from Table 8. This implies there is no different impact of test length for ethnicity group composition in U.S. studies.

Table 8: Regressions of the standardized gender difference on the two respective test lengths, a dummy for a primarily minority group sample and an interaction term. Observations are weighted and standard errors are clustered by article number.

Dependent variable: standardized gender difference	**Number of questions**	**Maximum amount of minutes**
Sample size	68	48
R-squared	0.2100	0.2234
Test length measure	-.0023*** (.0008)	.00145** (.0006)
Primarily minority group dummy	-.2699** (.1171)	-.2869* (.1469)
Test length * primarily minority group dummy	.0020 (.0020)	.0021 (.0018)

5.2.1.4 Accounting for age

The findings from Table 9 imply different age groups respond differently to test lengths. In general, the explanatory power of the combination of only a limited amount of variables is noteworthy, as indicated by a large R-squared. The missing observations seem used as the base value in the regression with the number of questions as our independent variable of interest. For the interaction terms, the adult/general population is used as the baseline value. In the regression using the maximum amount of minutes, this is the elementary school dummy. What is noteworthy is that nearly all coefficient estimates are significantly. What is especially clear is that for one length measure all interaction terms are significantly negative, for the other they are positive. Given that these direction are the other way around for both the test length measure and the age dummy, this helps explain otherwise contradicting findings. Because of this and the somewhat similar sizes of the estimated significant parameters, I am cautious to conclude there is a different impact of test length on the math gender gap based on sample age.

Table 9 regressions of the standardized gender difference on the two respective test lengths, dummies for the different age groups and their interaction term. Observations are weighted and standard errors are clustered by article number.

Dependent variable: standardized gender difference	**Number of questions**	**Maximum amount of minutes**
Sample size	282	163
R-squared	0.610	0.638
Test length measure	.0045** (.0017)	-.007*** (.0012)
Elementary school dummy	.6419*** (.1611)	(omitted)
Middle school dummy	.7536*** (.1604)	-.2786** (.1188)
High school dummy	.9523*** (.1080)	-.2311 (.1628)
College dummy	.9908*** (.0927)	-.0427 (.1181)
Adult/general population dummy	.3225*** (.0345)	-.8679*** (.2367)
Test length * Elementary school dummy	-.0035 (.0029)	(omitted)
Test length * Middle school dummy	-.0059** (.0025)	.0037** (.0018)
Test length * High school dummy	-.0053** (.0021)	.0082***(.0017)
Test length * College dummy	-.0051** (.0020)	.0061*** (.0020)
Test length * Adult/general population dummy	(omitted)	.0171 (.0112)

5.2.2 Accounting for timed vs. untimed tests

In the regression estimate of subsection 5.1 the variable of a test being timed or not had only been included as a control variable. Three dummies were included for this: if it was timed, if it was untimed or if it was unknown. An imposed time limit will, however, directly impact the experienced test length. Hence, I here consider whether a study had a time limit (so only including these observations for which we know there was a maximum time imposed), did not have a time limit, it was unknown if there was a time limit or if it was either unknown or timed. As there were 50 items for which the amount of questions was known but not whether these tests were timed or not, at least part may be timed and it could be interesting to include these. The results are displayed in Table 10.

No significant impact on the math gender gap is found when only using timed tests. The missing data subsample yields a negative coefficient (at the 1%) level of -.0035 when the test length measure is the number of questions. Also the combined subset of observations which were timed or for which whether they are timed or not is unknown yields a significant estimate of -.0053, at the 1% level for this measure. Ergo: the math gender gap increases in favour of girls as additional questions are added to a test for these groups – which is in line with the hypothesis of a female favoured test length bias.

We find significant findings for untimed tests with the minute maximum as test length measure, but this is based on only three observations, so negligible. We find no significant impact of either test length measure when only including timed tests.

Table 10: Simple linear regression for subsamples based on whether they were 1) timed, 2) untimed, 3) it was unknown if they were missing or 4) timed and missing. Standard errors are clustered by article. Article 87 is excluded.

Dependent variable: standardized gender difference	**Number of questions**		**Maximum amount of minutes**	
	weighted	unweighted	weighted	unweighted
Timed tests	.0017 (.0025)	-.0016 (.0011)	-.0032 (.0021)	-.0008 (.0010)
Sample size	189	202	161	174
Untimed tests	-.0016 (.0015)	.0017 (.0030)	-.0563*** (.0000)	
Sample size	45	45	3	3
Missing data	-.0035*** (.0012)	.0005 (.0013)	Too few observations	Too few observations
Sample size	50	50	-	-
Timed and missing observations	.0053** (.0022)	-.0011 (.0009)	.0023 (.0016)	-.0007 (.0010)
Sample size	240	253	163	176

5.2.3 Accounting for missing variables

As was discussed previously in section 4.2 and 4.3 and in the overall regression estimate, there are some missing variables. In order to consider the impact of those missing, I compare the estimates of a full weighted regression model using only the known values to the outcomes where the missing variables are set as a dummy (as in subsection 5.1). For the maximum amount of minutes variable we find an estimated coefficient of -.0165 (.0236 standard error, N= 25) and for the number of questions .0014 (.00065 se, N=35), the latter of which is significantly different from zero at a 5% level. As the direction of the estimate is positive in this case, this may indicate non-random missing variable observations. This was to be expected given the significant differences between observed and unobserved control variables in section 4.3.

5.2.4 Accounting for world regions

As described in section 2.1.1, the literature indicates that the gender gap is related to the sample's nationality. Although Lindberg et al. (2010) did not find a significant impact of world region on the math gender gap – conceivably for being world regions rather than countries - it proves of interest to consider how different world regions respond differently to test lengths. A significant outcome is found when isolating the European subpopulation and using either test length measures, as can be seen from table 11. A coefficient of -0.010 is here estimated: for every 10 additional minutes, the standardized math gender gap is estimated to decrease by 0.10. The corresponding t-value is -3.94, which indicates a highly significant corresponding p-value below 0.001. Given the constant term of 0.40 in the European regression, a test would need to consist of 40 minutes to eliminate the math gender gap in favour of boys. The -0.008 estimated coefficient for the number of questions parameter in the European subsample and a constant term of .38 indicates a comparable situation.

Particularly the findings from Europe are noteworthy due to their high significant levels. When including an interaction term in European samples, both test length and the world region's dummy variable are significantly different from zero at the 1% level. The interaction effect is the only one pointing into a negative direction in such a regression though.

Table 11 Simple linear regression per world region. Impact of test length in number of questions and maximum minutes on d-value (only included are those regressions where n>20), standard errors are clustered by article number and weighted values are used, article 87 is excluded.

Dependent variable: standardized difference	**USA**	**Europe**	**Asia**
Maximum minutes – R-squared	0.0004	0.54	0.013
N	96	28	31
Estimated impact on standardized difference per minute added	.0001 (.0007)	-.0102 *** (.003)	.0012 (.0033
Number of questions – R-squared	0.0188	0.6022	-
N	164	42	-
Estimated impact on standardized difference per question added	-.0013 (.0010)	-.0078*** (.0019)	-

From a visual representation of the European subsample in Figure 26 and Figure 27 it can be observed that this significant effect is in part driven by the two large sample sizes of studies with relatively long tests and where girls obtained higher scores than boys. Combined with the limited sample sizes of 28 and 42, this again indicates the need for further research.

Figure 26: weighted scatterplot and fitted regression line of European samples, regression of the standardized difference on the number of questions

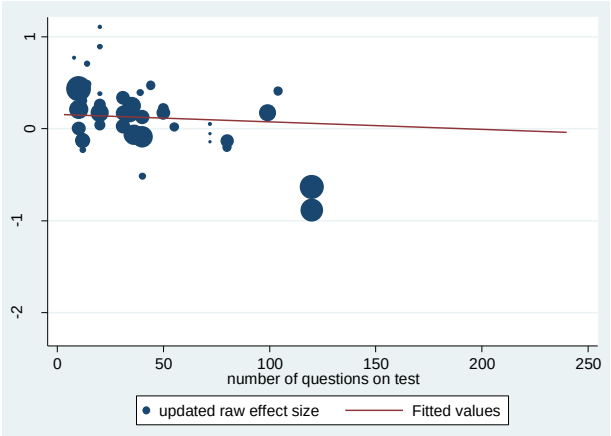
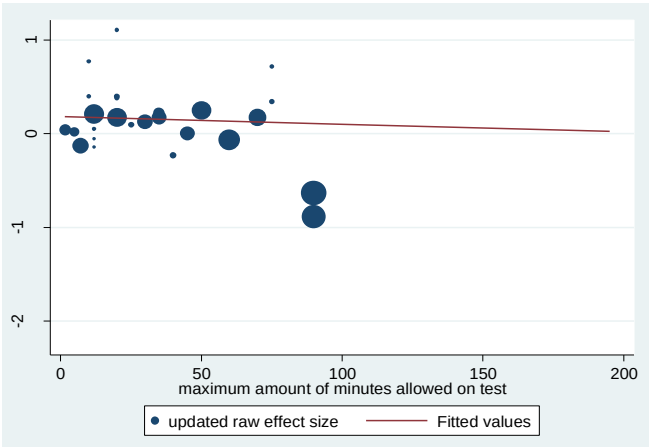


Figure 27: weighted scatterplot and fitted regression line of European samples, regression of the standardized difference on the maximum amount of minutes available in observations.



5.2.5 Accounting for polynomials

As elaborated on in subsection 4.1.3, the regression estimates could be improved by adding polynomials of the length measures to the regressions. There may, for example, be a particular length where girls or boys benefit more, after which the effect deteriorates. Such would not be identified by a linear term only, but could be captured by polynomial terms. For reference: the estimates of the test length effect on the math gender gap without polynomials (so the first regression in subsection 5.1.) was -.0008 and insignificant for both measures of test length. Including squared and cubic terms add predictive power to the regression (see Table 12). The adding of a squared term to the maximum amount of minute regression at first seems of some value – though the squared term's size being <.0000 indicates little practical relevance of this term in itself. It is moreover observable the regressions yield significant impact particularly in the case of cubic terms with weighted observations for both test length measures. However, the theoretical added value of a cubic terms in this situation seems limited. This is also observable from the <.0000 estimated coefficient, as it would imply multiple kinks in the relation between test length and gender gap, which simply seems unlikely. All in all, I do not consider these findings of much interst.

Table 12: including polynomials of test length. (1) and (4) relate to adding squared test length, (2) and (5) to adding cubed and (3) and (6) adds weights to the cubed regressions. Standard errors clustered by article number. ° symbolizes an R-squared is unadjusted.

Dependent variable: standardized gender difference	**Maximum amount of minutes**			**Number of questions**		
	(1)	(2)	(3)	(4)	(5)	(6)
Sample size	179	179	166	298	298	285
Adjusted R-squared	.0346	0.0370	0.1544°	-.0005	0.0127	0.3899°
Test length	-.0059** (.0026)	-.0009 (.0058)	.0266** (.0110)	.0001 (.0020)	.0069 (.0050)	.0275*** (.0089)
Squared test length	.0000** (.0000)	-.0001 (.0001)	-.0004** (.0002)	-.0000 (.0000)	-.0001* (.0001)	-.0003** (.0001)
Cubed test length	-	-.0000 (.0000)	-.0000** (.0000)	-	-.0000* (.0000)	-.0000** (.0000)

6 Conclusion & discussion

6.1 Conclusions

This thesis contributes to the literature by providing a first investigation in the potential effect of test length on the math gender gap. The findings from the extended meta-analysis by Lindberg et al. (2010) are mixed though at times suggest there is a small negative impact of test length on the math gender gap. This is of relevance due to individual economic opportunity of either gender and nation's impact of mathematics skills. Both require a just assessment of math ability and progress. First: individuals may be valued higher in the labour market when having (advanced) math skills. Second: national policies in various countries focus on an increase in the labour market share in STEM fields, as this could spark innovation.

In line with the findings from Balart & Oosterveen (2017), the results from this thesis indicate that test length may induce a bias threat for tests as an evaluation mechanism. The regression that includes all controls and known weights indicates a significant decrease of the standardized gender gap of .002 per added question. The comparable regression where the maximum minutes allowed is used as a test length measure yields a similar coefficient, though nonsignificantly different from zero. Both test type, ability levels, ethnicity of U.S. groups and age indicate potential interaction effects for these categories of participants and tests. Limiting the dataset to timed values generates insignificant effects of increased test length on the math gender gap. However, significantly negative coefficients for the subgroups for which it was unknown if observations were timed or not are estimated. The same is true when including both this group and the timed studies in a regression. The European subsample stands out by indicating a highly significant impact on the gender gap per added minute (of -.010) and question (-.008).

More controversial is that a significantly positive impact of question amount on the gender was found when weights were estimated and added where they were missing. Also a positive significant estimate when no dummies are used so as to include missing variables stresses the mixed nature of the results. What is important in this is the limited internal validity caused by the missing variables. The oftentimes significantly different characteristics of short and long tests, as discussed in subsection 4.4, lead to a cautionary note. There are more primarily minority samples in longer tests and there is an overrepresentation general population in longer tests. Both are related to lower standardized gender gaps and may prove to be confounding variables that may not be fully accounted for in the regressions.

The combination of findings highlights the mixed nature of the results. At this point, no conclusive statement on the existence of a female favoured test bias as length increases is justified. However, given that the vast majority of the regressions that are estimated do hint at a negative relationship (though oftentimes insignificantly), the findings are promising enough to justify and encourage future research into the topic. For broader consideration is that the *within* test performance decline may be larger for males than for females, even if it does not show when tests get longer. In any exam, regardless of its length, there may be a performance decline that is stronger for boys on average. This could still exist, and is indeed implied by the research of Balart & Oosterveen (2017), regardless of the findings in this thesis.

It is particularly worthwhile, given the many limitations of the set-up of this thesis as will be elaborated on the coming subsection, to work with other research strategies.

6.2 Limitations & suggestions for further research

Although this set-up seems justified at the current exploratory level into a new hypothesis, its set-up has some inherent limitations. First of all, this is due to the research being based on secondary data. A consideration here is that 'not all research is created equal'. Limitations within all 441 observations are limitations within this study. If there was (selection)bias, a flaw in the set-up, measurement errors, or any other issue it will work through in the final results and I have no insight in this. Weighing the articles on the standard errors does not do justice to all quality differences – although it does serve as an indication and is the best available. A second limitation to the use of secondary data, is that oranges are compared to apples. There is likely variation in the studies beyond what is controlled for, this is of particular note in the two test length measures. In reality not every test requires the time that is set for it, nor is every question by definition the same length. What was counted as one question could differ per study - every subquestion, every full question. As I primarily relied on author's answers and what was mentioned in the articles, this measurement error is a large concern. A positive aspect of the study however is that two test length measures are used, rather than one, which should in part account for this.

Second, despite the careful consideration of the authors that shared the primary dataset with us, not all values could be checked again (i.e. test design, level, etc.). As there were some alterations made to the original dataset for the variables that I constantly checked (such as means and standard deviations - these alterations were expanded upon in subsection 3.2), there may well be other (notation) errors from the original, or updated, dataset that may have led to large or small bias in the estimate. It is however possible that these work in both directions and cancel each other out.

Third, when we combine our dataset particularly with the notion elaborated on in subsection 2.2.1 on variability, the limited external validity is to be considered. Hedges & Nowell (1995) indicate boys perform better than girls by a small amount at best, but the authors find substantially more boys in the high performing groups. Halpern & Benbow (2007) conclude the observed male advantage in mathematics is largest at the upper end of the ability distribution – also the economically most relevant part. Considering that not a single study in the used dataset includes questions from the tranche of the highest depth of knowledge, level 4, this could be a reason that only a small gender gap is found. This may harm the wider applicability of our test length impact estimate particularly for the region that is economically and equitably the most interesting.

Fourth, one ought to consider that preselection may have taken place in groups studied. Considering that a substantial amount of studies are held under group of students that are already part of another selection, for instance certain study fields at universities or high school levels, pre-selection may have already taken place. Psychology students are, regardless of gender, likely to have math skills in a given range – plus exceptions of course. This study, as these skills, may correlate with a certain gender ratio. Those with a poor level of statistics will not have chosen the study, so will by definition not be in the sample group. Those with very high math skill level may also not be in psychology, as they have chosen studies more in line with such a skillset. If it is the case that men/women are due to these underlying math skills into a certain sample studied in the first place, comparing across genders within one group that is implicitly selected based on math will imply a smaller gender gap than actually exists. Smith & White's (2001) article provides a good example of an explicit occurrence of this. Here men and women are in advance clustered into high & low identity groups. Women are hereby overrepresented in low

identified group, already indicating lower achievement but next the comparisons are made within these tranches. Ergo, the dataset used likely understates the math gender gap.

Fifth, specific to our dataset, Lindberg et al. (2010) have, justifiably for their study, at times chosen to merge several test results of participants into one observation. In hindsight a limitation of the format of this thesis was that it did not allow for adding new studies and splitting them in accordance with the interest of this hypothesis. It may be of value in future to account for this.

Sixth, that the studies for which test length was available estimated a significantly higher standardized gender gap than the studies that were not included due to a lack of information of their test length is potentially problematic. Both for reasons of selection bias and external validity, as was elaborated on in section 4.3. Also the harmed homoscedasticity of the error terms for the more extended models and potential added effects of introducing squared and cubic terms could limit the value of the estimates. The different characteristics of short versus long tests has already been mentioned in the conclusion as a potential limitation of the internal validity.

Promising further research in the hypothesis of female favoured test bias as math test length increases may come in two forms. An ideal experiment for the hypothesis would be a randomized trial, with different randomized comparable groups making the same difficulty of tests with the only variable of difference being the tests' length. A significantly smaller male favored gender gap in the treatment group as compared to the control group would indicate the existence of test length impact. Second, other research that has already been done could be exploited where by coincidence all else was kept equal, except for test duration – perhaps PISA, CITO or TIMMS tests changed their length some point in time. From an economic perspective, particularly (nation-wide) high-stakes tests would be of interest, given their impact on career perspectives.

6.3 Policy implications

The main lesson to take away from both the literature review and the data analysis is the impact that test design could have on outcomes of different groups being tested. An examiner or test-maker should be aware that certain factors of assessment, including length, could favour one group over the other. As the literature review indicates, fairness and an unbiased setting in which both genders perform optimal does go beyond design – stereotypes, pressure and anxiety levels may well be of stronger impact.

Policy implications to tackle test length impact will only become relevant once more conclusive evidence becomes available. Hypothesizing that there is indeed a female favoured bias as test length increases, it would be of importance to research why it is there. The underlying rationale of a potential larger performance decline for boys is unknown. One hypothesis is that girls hold more intrinsic motivation and hence want to do well no matter what (be this an innate or taught skill), regardless of stake or required endurance. Second, one could argue that test anxiety diminishes as time progresses, so the longer someone has, the more time one would have to calm down and ensure the knowledge is accessed. In the latter, the female advantage of longer tests stems from overcoming their hurdles as compared to males. If this is the way in which test length impacts the gender gap, rather than speaking of female favoured bias, it could also diminish male favoured bias – not of design but of a gender's tendency. Evidently, a test is generally considered to be a valid test when the actual skill (math in this case) is tested. Not when factors such as stamina, needing time to reduce your math anxiety, and concentration help or hinder you from obtaining a score in line with your skills. Ergo, the reason for this difference is important before one can justify a certain length of test has the least gender bias.

7 Bibliographies

7.1 Thesis bibliography

Azmat, G., Calsamiglia, C., & Iriberri, N. (2016). Under pressure: Gender differences in exam stakes. Retrieved October 10, 2016, from http://voxeu.org/article/how-school-children-respond-exam-pressure

Baker, D. P., & Jones, D. P. (1993). Creating gender equality: Cross-national gender stratification and mathematical performance. *Social Education*, *66*, 91–103.

Balart, P., & Oosterveen, M. (2017). *Gender Differences in Testing Behaviour*.

Baldiga, K. (2013). Gender differences in willingness to guess. *Management Science*. Retrieved from http://pubsonline.informs.org/doi/abs/10.1287/mnsc.2013.1776

Bandura, A. (1994). *Self-efficacy*. John Wiley & Sons.

Bastalich, W., & Mills, J. (2003). "I Had This Real Feeling That It Was a Boys Club." … *Education for a …*. Retrieved from https://search.informit.com.au/documentSummary;dn=935933553046890;res=IELENG

Bedard, K., & Cho, I. (2009). Early Gender Test Score Gaps Across OECD Countries. *Economics of Education Review*, *29(3)*, 348–363.

Beede, D., & Julian, T. (2011). Women in STEM: A gender gap to innovation. *Economics and …*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1964782

Beller, M., & Gafni, N. (2000). Can Item Format (Multiple Choice vs. Open-Ended) Account for Gender Differences in Mathematics Achievement? *Sex Roles*, *42*(1), 1–21.

Buonanno, P., & Pozzoli, D. (2009). Early labour market returns to college subject. *Labour*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9914.2009.00466.x/full

Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, Competitiveness, and Career Choices. *Quarterly Journal of Economics*, *129*(3), 1409–1447. https://doi.org/10.1093/qje/qju009

Caprile, M., Palmén, R., Sanz, P., & Dente, G. (2015). *Encouraging STEM studies for the labour market*.

CBS. (2016). *The Netherlands on the European Scale 2016*.

Cheema, J. R., & Galluzzo, G. (2013). Analyzing the Gender Gap in Math Achievement: Evidence from a Large-Scale US Sample. *Research in Education*, *90*(1), 98–112.

Close, S., & Shiel, G. (2009). Gender and PISA Mathematics : Irish results in context, *8*(1), 20–33.

Correll, S. J. (2001). Gender and the Career Choice Process: The Role of Biased Self-Assessments on JSTOR. *American Journal of Sociology*, *106*(6), 1691–1730. Retrieved from http://www.jstor.org/stable/10.1086/321299

Correll, S. J. (2004). Constraints into Preferences: Gender, Status, and Emerging Career Aspirations. *American Sociological Review*, *69*(1), 93–113. https://doi.org/10.1177/000312240406900106

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*. Retrieved from http://www.ingentaconnect.com/content/aea/jel/2009/00000047/00000002/art00003

de Lange, J. (1987). *Mathematics: insights and meaning*. OC&OW.

Dubner, S. J. (2016a). The True Story of the Gender Pay Gap. Retrieved from

http://freakonomics.com/podcast/the-true-story-of-the-gender-pay-gap-a-new-freakonomics-radio-podcast/

Dubner, S. J. (2016b). What Are Gender Barriers Made Of? Retrieved from http://freakonomics.com/podcast/gender-barriers/

Else-quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-National Patterns of Gender Differences in Mathematics : A Meta-Analysis, *136*(1), 103–127. https://doi.org/10.1037/a0018053

Ernesto, R., Paola, S., & Luigi, Z. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(12), 4403–4408.

Espinosa, M., & Gardeazabal, J. (2013). Do students behave rationally in multiple choice tests? Evidence from a field experiment. *Journal of Economics and …*. Retrieved from https://ideas.repec.org/a/jec/journl/v9y2013i2p107-135.html

European Commission. (2015). *Education and Training Monitor 2015*.

Forgasz, H. (2006). Australian Year 12 Mathematics Enrolments: Patterns and Trends-Past and Present. Retrieved from http://arrow.monash.edu.au/hdl/1959.1/205482

Forsthuber, B., Horvath, A., & Motiejunaite, A. (2010). *Gender differences in educational outcomes: Study on the measures taken and the current situation in Europe.*

Fryer, R. G., & Levitt, S. D. (2009). An Empirical Analysis of the Gender Gap in Mathematics.

Giulia, F., Silvia, G., Francesca, C., & Caterina, P. (2014). Implicit gender–math stereotype and women's susceptibility to stereotype threat and stereotype lift. *Learning and Individual Differences*, *32*, 273–277.

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *QUARTERLY JOURNAL OF …*. Retrieved from http://www.nber.org/~rosenbla/econ311-05/syllabus/murielgender.pdf

Gneezy, U., & Rustichini, A. (2004). Gender and competition at a young age. *The American Economic Review*. Retrieved from http://www.jstor.org/stable/3592914

Grogger, J., & Eide, E. (1995). Changes in college skills and the rise in the college wage premium. *Journal of Human Resources*. Retrieved from http://www.jstor.org/stable/146120

Grossman, G., & Helpman, E. (1991). R & D Spillovers and the Geography of Innovation and Production. *Production*, *86*(3), 630–640. https://doi.org/Article

Halpern, D., & Benbow, C. (2007). The science of sex differences in science and mathematics. … *Science in the Public …*. Retrieved from http://psi.sagepub.com/content/8/1/1.short

Hannabus, K. C. (1991). Mixed results. *Oxford Magazine*, *74*, 4–5.

Hannabus, K. C. (1992). Mixed results. *The Cambridge Review*, *113*, 40–42.

Hanushek, E., & Woessmann, L. (2008). Education and economic growth. *Education …*. Retrieved from http://search.proquest.com/openview/30af4349f3470351826d056fd6a6c969/1?pq-origsite=gscholar&cbl=1766362

Hedges, L. V, & Nowell, A. (1995). Sex Differences in Mental Test Scores , Variability , and Numbers of High-Scoring Individuals, *269*(July).

Held, J., Alderton, D., Foley, P., & Segall, D. (1993). Arithmetic reasoning gender differences –

Explanations found in the armed services vocational aptitude battery (ASVAB). *Learning and Individual Differences*, *5*(171–186).

Helme, S., & Lamb, S. (2007). Student experiences of VCE further mathematics. *… : Essential Research, Essential Practice. Proceedings of …*. Retrieved from http://files.eric.ed.gov/fulltext/ED503746.pdf#page=359

Hill, C., Corbett, C., & Rose, A. S. (2010). Why so few? Women in Science, Technology, Engineering, and Mathematics. Retrieved from http://eric.ed.gov/?id=ED509653

Hutt, C. (1972). *Males and Females*.

Hyde, J., & Mertz, J. (2009). Gender, culture, and mathematics performance. *… of the National Academy of Sciences*. Retrieved from http://www.pnas.org/content/106/22/8801.full.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological Bulletin*, *107*(2), 139–155.

Joensen, J., & Nielsen, H. (2010). More Successful because of Math: Combining a Natural Experiment and a Structural Dynamic Model to Explore the Underlying Channels. *Unpublished Working Paper*. Retrieved from https://www.economicdynamics.org/meetpapers/2011/paper_995.pdf

Johnson, A. C. (2007). Unintended Consequences: How Science Professors Discourage Women of Color. *Science Education*, *91*(1), 36–74. https://doi.org/10.1002/sce

Koedel, C., & Tyhurst, E. (2012). Math skills and labor-market outcomes: Evidence from a resume-based field experiment. *Economics of Education Review*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0272775711001531

Legewie, J., & DiPrete, T. A. (2014). The High School Environment and the Gender Gap in Science and Engineering. *Sociology of Education*, *87*(4), 259–280.

Lindberg, S. M., Hyde, J. S., & Hirsch, L. M. (2008). Gender and Mother-Child Interactions during Mathematics Homework: The Importance of Individual Differences. *Merrill-Palmer Quarterly*, *54*(2), 232–255.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: a meta-analysis. *Psychological Bulletin*, *136*(6), 1123–1135. https://doi.org/10.1037/a0021276.New

Maccoby, E. E. (1966). *The Development of Sex Differences*. Stanford University Press.

McCrum, N. G. (1991). A fair admissions system. *Oxford Magazine*, *72*, 16–17.

Mehrens, W., Millman, J., & Sackett, P. (1994). Accommodations for candidates with disabilities. *The Bar Examiner*. Retrieved from http://scholar.google.nl/scholar?q=Mehrens%2C+Millman%2C+and+Sackett+%281994%29+&btnG=&hl=nl&as_sdt=0%2C5#0

Mellors-Bourne, R., Connor, H., & Jackson, C. (2011). *Stem graduates in non-STEM jobs: Executive Summary*.

Microsoft. (2016). *Why Europe's girls aren't studying STEM*.

National Science Foundation. (2016). *Science & Engineering Indicators 2016*. National Science Board. https://doi.org/10.1002/ejoc.201200111

Noailly, J., Waagmeester, D., Jacobs, B., Rensman, M., & Webbink, D. (2005). *Scarcity of science and engineering students in the Netherlands*. The Hague. Retrieved from

https://www.cpb.nl/sites/default/files/publicaties/download/scarcity-science-and-engineering-students-netherlands.pdf

Nollenberger, Natalia ; Rodríguez-Planas, Núria; Sevilla, A. (2016). The Math Gender Gap: The Role of Culture. *The American Economic Review*, *106*(5), 257–261.

OECD, O. for E. C. and D. (2010). PISA 2009 results: what students know and can do: student performance in reading, mathematics and science (volume I). *OECD*, *I*.

OECD Organisation for Economic Co-operation and Development. (2004). Learning for Tomorrow's World - First results from pisa 2003. *OECD Publications*. Retrieved from https://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/34002216.pdf

OECD Organisation for Economic Co-operation and Development. (2014). *PISA 2012 Results in Focus*.

Olson, S., & Riordan, D. G. (2012). *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. Washington, DC.

Ors, E., Palomino, F., & Peyrache, E. (2013). Performance gender gap: does competition matter? *Journal of Labor Economics*. Retrieved from http://www.jstor.org/stable/10.1086/669331

Paglin, M., & Rufolo, A. (1990). Heterogeneous human capital, occupational choice, and male-female earnings differences. *Journal of Labor Economics*. Retrieved from http://www.jstor.org/stable/2535301

Pekkarinen, T. (2012). *Gender Differences in Education*. *Nordic Economic Policy Review*.

Pekkarinen, T. (2015). Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations. *Journal of Economic Behavior & Organization*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167268114002261

Pope, D. G., & Sydnor, J. R. (2010). Geographic Variation in the Gender Differences in Test Scores. *Journal of Economic Perspectives*, *24*(2), 95–108.

Reardon, S., Fahle, E., Kalogrides, D., Podolsky, A., & Zarate, R. (2016). Test Format and the Variation of Gender Achievement Gaps within the United States. *Social Cognition*, *34*(3), 196–216.

Romer, P. M. (1990). Endogenous Technological Change. *Journal of Political Economy*, *98*(5), S71–S102. https://doi.org/10.1086/261725

Rose, H., & Betts, J. (2001). Math matters: The links between high school curriculum, college graduation, and earnings. Retrieved from http://scholar.google.nl/scholar?hl=nl&q=Rose+and+Betts+%282001&btnG=&lr=#0

Rose, H., & Betts, J. (2004). The effect of high school courses on earnings. *Review of Economics and Statistics*. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/003465304323031076

Schrøter Joensen, J., & Skyt Nielsen, H. (2015). Mathematics and Gender: Heterogeneity in Causes and Consequences. *The Economic Journal*, *126*(593), 1129–1163.

Scott, E. C., Page, M. E., & West, J. E. (2010). Sex and Science: How Professor Gender Perpetuates the Gender Gap. *Quarterly Journal of Economics*, *125*(3), 1101–1144.

Shields, S. (1982). The variability hypothesis: The history of a biological model of sex differences in intelligence. *Signs*, *7*(4), 769–797. Retrieved from http://www.jstor.org/stable/3173639

Smith, J. L., & White, P. H. (2001). Development of the domain identification measure: A tool for investigating stereotype threat effects. *Educational and Psychological Measurement*, *61*, 1040–1057.

Snyder, T. D., Dillow, S. A., & Hoffman, C. H. (2008). Digest of Education Statistics, 2007. *National Center for Education Statistics*.

Steffens, M. C., Jelenec, P., & Noack, P. (2010). On the leaky math pipeline: Comparing implicit math-gender stereotypes and math withdrawal in female and male children and adolescents. *Journal of Educational Psychology*, *102*(4), 947–963.

Tannenbaum, D. (2012). Do gender differences in risk aversion explain the gender gap in SAT scores? Uncovering risk attitudes and the test score gap. *Unpublished Paper, University of Chicago, Chicago*. Retrieved from http://scholar.google.nl/scholar?q=Tannenbaum+2012+gender&btnG=&hl=nl&as_sdt=0%2C5#0

Tiedemann, J. (2000). Parents' gender stereotypes and teachers' beliefs as predictors of children's concept of their mathematical ability in elementary school. *Journal of Educational Psychology*, *92*(1), 144.

TU Delft. (2015). FAQ BSc. Technische Wiskunde. Retrieved November 20, 2016, from http://www.tudelft.nl/fileadmin/Files/tudelft/studeren/bachelor/FAQ_nw/FAQ_BSc_Technische_Wiskunde.pdf

U.S. Department of Education. (2015). Science, Technology, Engineering and Math: Education for Global Leadership. Retrieved March 20, 2017, from https://www.ed.gov/Stem

Watt, H. (2005). Explaining gendered math enrollments for NSW Australian secondary school students. *New Directions for Child and Adolescent …*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/cd.147/abstract

Zoller, U., & Ben-Chaim, D. (1989). Gender differences in examination type preferencees, test anxiety, and academic achievements in college science education. In *Fifth GASAT Conference*. Haifa, Israel.

7.2 Meta-analysis bibliography

1. Abdel-Khalek AM, Lynn R. Sex differences on the standard progressive matrices and in educational attainment in Kuwait. Personality and Individual Differences. 2006;40:175–182.
2. Abedi J, Lord C. The language factor in mathematics tests. Applied Measurement in Education.2001;14:219–234.
3. Akerman BA. Twins at puberty: A follow-up study of 32 twin pairs. Psychology: The Journal of the Hellenic Psychological Society. 2003;10:228–236.
4. Alkhateeb HM. Gender differences in mathematics achievement among high school students in the united arab emirates, 1991–2000. School Science and Mathematics. 2001;101:5–9.
5. Alkhateeb HM. A preliminary study of achievement, attitudes toward success in mathematics, and mathematics anxiety with technology-based instruction in brief calculus. Psychological Reports.2002;90:47–57. [PubMed]
6. Alkhateeb HM, Jumaa M. Cooperative learning and algebra performance of eighth grade students in united arab emirates. Psychological Reports. 2002;90:91–100. [PubMed]

7. Anderman EM, Midgley C. Changes in achievement goal orientations, perceived academic competence, and grades across the transition to middle-level schools. Contemporary Educational Psychology. 1997;22:269–298. [PubMed]
8. Arigbabu AA, Mji A. Is gender a factor in mathematics performance among Nigerian preservice teachers? Sex Roles. 2004;51:749–753.
9. Atkins M, Rohrbeck CA. Gender effects in self-management training: Individual cooperative interventions. Psychology in the Schools. 1993;30:362–368.
10. Aunio P, Hautamäki J, Heiskari P, Van Luit JEH. The early numeracy test in Finnish: Children's norms. Scandinavian Journal of Psychology. 2006;47:369–378. [PubMed]
11. Austin JT, Hanisch KA. Occupational attainment as a function of abilities and interests: A longitudinal analysis using project TALENT data. Journal of Applied Psychology. 1990;75:77–86.[PubMed]
12. Badian NA. Persistent arithmetic, reading, or arithmetic and reading disability. Annals of Dyslexia.1999;49:45–70.
13. Bandalos DL, Yates K, ThorndikeChrist T. Effects of math self-concept, perceived self-efficacy, and attributions for failure and success on test anxiety. Journal of Educational Psychology. 1995;87:611–623.
14. Battista MT. Spatial visualization and gender differences in high school geometry. Journal for Research in Mathematics Education. 1990;21:47–60.
15. Baumert J, Demmrich A. Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. European Journal of Psychology of Education.2001;16:441–462.
16. Bell SM, Mccallum RS, Bryles J, Driesler K, Mcdonald J, Park SH, Williams A. Attributions for academic-success and failure : An individual difference investigation of academic-achievement and gender. Journal of Psychoeducational Assessment. 1994;12:4–13.
17. Bempechat J, Graham SE, Jimenez NV. The socialization of achievement in poor and minority students - A comparative study. Journal of Cross-Cultural Psychology. 1999;30:139–158.
18. Bennett RE, Morley M, Quardt D, Rock DA. Graphical modeling: A new response type for measuring the qualitative component of mathematical reasoning. Applied Measurement in Education. 2000;13:303–322.
19. Benson J, Bandalos D, Hutchinson S. Modeling test anxiety among men and women. Anxiety, Stress, and Coping. 1994;7:131–148.
20. Bibby PA, Lamb SJ, Leyden G, Wood D. Season of birth and gender effects in children attending moderate learning difficulty schools. British Journal of Educational Psychology. 1996;66:159–168.[PubMed]
21. Bielinski J, Davison ML. Gender differences by item difficulty interactions in multiple-choice mathematics items. American Educational Research Journal. 1998;35:455–476.
22. Birenbaum M, Gutvirtz Y. The relationship between test anxiety and seriousness of errors in algebra. Journal of Psychoeducational Assessment. 1993;11:12–19.
23. Birenbaum M, Nasser F. Ethnic and gender differences in mathematics achievement and in dispositions towards the study of mathematics. Learning and Instruction. 2006;16:26–40.
24. Birenbaum M, et al. Stimulus features and sex differences in mental rotation test performance. Intelligence. 1994;19:51.
25. Bliwise NG. Web-based tutorials for teaching introductory statistics. Journal of Educational Computing Research. 2005;33:309.
26. Bolger N, Kellaghan T. Method of measurement and gender differences in scholastic achievement. Journal of Educational Measurement. 1990;27:165–174.

27. Borg MG. Sex and age differences in the scholastic attainment of grammar school children in the first three years of secondary schooling: A longitudinal study. Research in Education. 1996;(56):1–20.
28. Borg MG, Falzon JM. Birth date and sex effects on the scholastic attainment of primary schoolchildren: A cross-sectional study. British Educational Research Journal. 1995;21:61.
29. Borg MG, Falzon JM, Sammut A. Age and sex differences in performance in an 11-plus selective examination. Educational Psychology. 1995;15:433–443.
30. Bornholt LJ, Goodnow JJ, Cooney GH. Influences of gender stereotypes on adolescents perception of their own achievement. American Educational Research Journal. 1994;31:675–692.
31. Brennan RT, Kim J, Wenz-Gross M, Siperstein GN. The relative equitability of high-stakes testing versus teacher-assigned grades: An analysis of the Massachusetts comprehensive assessment system (MCAS) Harvard Educational Review. 2001;71:173–216.
32. Bridgeman B, Harvey A, Braswell J. Effects of calculator use on scores on a test of mathematical reasoning. Journal of Educational Measurement. 1995;32:323–340.
33. Bridgeman B, Wendler C. Gender differences in predictors of college mathematics performance and in college mathematics course grades. Journal of Educational Psychology. 1991;83:275–284.
34. Bruce CK, Lawrenz FP. Actual and teacher perceptions of the abilities of mathematical high school chemistry students in Minnesota. School Science and Mathematics. 1991;91:1–5.
35. Bull R, Johnston RS. Children's arithmetical difficulties: Contributions from processing speed, item identification, and short-term memory. Journal of Experimental Child Psychology. 1997;65:1–24.[PubMed]
36. Busato VV, Dam G. T. M. t., Eeden P. v. d. Gender-related effects of co-operative learning in a mathematics curriculum for 12–16-year-olds. Journal of Curriculum Studies. 1995;27:667–686.
37. Byrnes JP, Hong L, Xing S. Gender differences on the math subtest of the scholastic aptitude test may be culture-specific. Educational Studies in Mathematics. 1997;34:49–66.
38. Byrnes JP, Takahira S. Explaining gender differences on SAT-math items. Developmental Psychology. 1993;29:805–810.
39. Byrnes JP, Takahira S. Why some students perform well and others perform poorly on SAT math items. Contemporary Educational Psychology. 1994;19:63–78.
40. Campbell E, Schellinger T, Beer J. Relationships among the ready or not parental checklist for school readiness, the Brigance Kindergarten and 1-grade screen, and SRA scores. Perceptual and Motor Skills. 1991;73:859–862.
41. Cankoy O, Tut MA. High-stakes testing and mathematics performance of fourth graders in north Cyprus. Journal of Educational Research. 2005;98:234–243.
42. Cardelle-Elawar M. Effects of feedback tailored to bilingual students' mathematics needs on verbal problem solving. Elementary School Journal. 1990;91:165–175.
43. Chan DW. Assessing giftedness of Chinese secondary students in Hong Kong: A multiple intelligences perspective. High Ability Studies. 2001;12:215–234.
44. Chen PP. Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. Learning and Individual Differences. 2002;14:77–90.
45. Cherian VI. Gender, socioeconomic-status, and mathematics achievement by Xhosa children. Psychological Reports. 1993;73:771–778.
46. Cherian VI, Cherian L. Relationship of divorce, gender, socioeconomic-status and unhappiness to mathematics achievement of children. Journal of Family Welfare. 1995;41:30–37.

47. Chipman SF, Marshall SP, Scott PA. Content effects on word problem performance: A possible source of test bias? American Educational Research Journal. 1991;28:897–915.
48. Clariana RB, Schultz CW. Gender by content achievement differences in computer-based instruction. The Journal of Computers in Mathematics and Science Teaching. 1993;12:277–288.
49. Collaer ML, Hill EM. Large sex difference in adolescents on a timed line judgment task: Attentional contributors and task relationship to mathematics. Perception. 2006;35:561–572. [PubMed]
50. Connors MA. Achievement and gender in computer-integrated calculus. Journal of Women and Minorities in Science and Engineering. 1995;2:113.
51. Crosser SL. Summer birth date children: Kindergarten entrance age and academic achievement. Journal of Educational Research. 1991;84:140.
52. Davies J, Brember I. Boys outperforming girls: An 8-year cross-sectional study of attainment and self-esteem in year 6. Educational Psychology. 1999;19:5–16.
53. Davis H, Carr M. Gender differences in mathematics strategy use - the influence of temperament. Learning and Individual Differences. 2001;13:83–95.
54. Davis-Dorsey J, Ross SM, Morrison GR. The role of rewording and context personalization in the solving of mathematical word problems. Journal of Educational Psychology. 1991;83:61–68.
55. De Brauwer J, Verguts T, Fias W. The representation of multiplication facts: Developmental changes in the problem size, five, and tie effects. Journal of Experimental Child Psychology. 2006;94:43–56.[PubMed]
56. De Lisle J, Smith P, Jules V. Which males or females are most at risk and on what? an analysis of gender differentials within the primary school system of Trinidad and Tobago. Educational Studies.2005;31:393–418.
57. DeMars CE. Gender differences in mathematics and science on a high school proficiency exam: The role of response format. Applied Measurement in Education. 1998;11:279–299.
58. Diamante T. Unitarian validation of a mathematical problem-solving exercise for sales occupations. Journal of Business and Psychology. 1993;7:383–401.
59. Dickhäuser O, Meyer W. Gender differences in young children's math ability attributions. Psychology Science. 2006;48:3–16.
60. Duffy J, Gunther G, Walters L. Gender and mathematical problem solving. Sex Roles. 1997;37:477–494.
61. Eid GK, Koushki PA. Secondary education programs in Kuwait: An evaluation study. Education (Chula Vista, Calif.) 2005;126:181–200.
62. El Hassan K. Gender issues in achievement in Lebanon. Social Behavior and Personality.2001;29:113–123.
63. Elliott JC. Affect and mathematics achievement of nontraditional college students. Journal for Research in Mathematics Education. 1990;21:160–165.
64. Entwisle DR, Alexander KL, Olson LS. The gender-gap in math – Its possible origins in neighborhood effects. American Sociological Review. 1994;59:822–838.
65. Evans EM, Schweingruber H, Stevenson HW. Gender differences in interest and knowledge acquisition: The United States, Taiwan, and Japan. Sex Roles. 2002;47:153–167.
66. Ewers CA, Wood NL. Sex and ability differences in childrens math self-efficacy and accuracy. Learning and Individual Differences. 1993;5:259–267.
67. Feldman R, Guttfreund D, Yerushalmi H. Parental care and intrusiveness as predictors of the abilities-achievement gap in adolescence. Journal of Child Psychology and Psychiatry and Allied Disciplines. 1998;39:721–730. [PubMed]

68. Fennema E, Carpenter TP, Jacobs VR. A longitudinal study of gender differences in young children's mathematical thinking. Educational Researcher. 1998;27:6–11.
69. Fink B, Brookes H, Neave N, Manning JT, Geary DC. Second to fourth digit ratio and numerical competence in children. Brain and Cognition. 2006;61:211–218. [PubMed]
70. Fischbein S. Biosocial influences on sex differences for ability and achievement test results as well as marks at school. Intelligence. 1990;14:127–139.
71. Fuller B, Hua H, Snyder C., Jr. When girls learn more than boys: The influence of time in school and pedagogy in Botswana. Comparative Education Review. 1994;38:347–376. [PubMed]
72. Gallagher AM, De Lisi R, Holst PC, McGillicuddy-De Lisi AV, Morely M, Cahalan C. Gender differences in advanced mathematical problem solving. Journal of Experimental Child Psychology.2000;75:165–190. [PubMed]
73. Galler JR, Ramsey FC, Harrison RH, Taylor J, Cumberbatch G, Forde V. Postpartum maternal moods and infant size predict performance on a national high school entrance examination. Journal of Child Psychology and Psychiatry. 2004;45:1064–1075. [PubMed]
74. Garner M, Engelhard G. Gender differences in performance on multiple-choice and constructed response mathematics items. Applied Measurement in Education. 1999;12:29–51.
75. Geary DC, Salthouse TA, Chen GP, Fan L. Are east Asian versus American differences in arithmetical ability a recent phenomenon? Developmental Psychology. 1996;32:254–262.
76. Geary DC, Saults SJ, Liu F, Hoard MK. Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. Journal of Experimental Child Psychology. 2000;77:337–353. [PubMed]
77. Gilbert WS. Bridging the gap between high school and college. Journal of American Indian Education. 2000;39:36.
78. Glutting JJ, Oh HJ, Ward T, Ward S. Possible criterion-related bias of the WISC-III with a referral sample. Journal of Psychoeducational Assessment. 2000;18:17–26.
79. Gottesman RL, Bennett RE, Nathan RG, Kelly MS. Inner-city adults with severe reading difficulties: A closer look. Journal of Learning Disabilities. 1996;29:589–597. [PubMed]
80. Gouchie C, Kimura D. The relationship between testosterone levels and cognitive ability patterns. Psychoneuroendocrinology. 1991;16:323–334. [PubMed]
81. Gresky DM, Eyck LLT, Lord CG, McIntyre RB. Effects of salient multiple identities on women's performance under mathematics stereotype threat. Sex Roles. 2005;53:703–716.
82. Grewal AS. Sex differences in algebra by senior secondary school students in Transkei, South Africa. Psychological Reports. 1998;83:1266–1266.
83. Halat E. Sex-related differences in the acquisition of the van hiele levels and motivation in learning geometry. Asia Pacific Education Review. 2006;7:173–183.
84. Hall CW, Davis NB, Bolen LM, Chia R. Gender and racial differences in mathematical performance. Journal of Social Psychology. 1999;139:677–689. [PubMed]
85. Hay I, Ashman AF, van Kraayenoord CE. The influence of gender, academic achievement and non-school factors upon pre-adolescent self-concept. Educational Psychology. 1998;18:461–470.
86. Hein J, Bzufka MW, Neumarker KJ. The specific disorder of arithmetic skills. prevalence studies in a rural and an urban population sample and their cliniconeuropsychological validation. European Child & Adolescent Psychiatry. 2000;9:87–101. [PubMed]
87. Held JD, Alderton DL, Foley PP, Segall DO. Arithmetic reasoning gender differences – Explanations found in the armed services vocational aptitude battery (ASVAB) Learning and Individual Differences. 1993;5:171–186.
88. Helwig R, Anderson L, Tindal G. Influence of elementary student gender on teachers' perceptions of mathematics achievement. Journal of Educational Research. 2001;95:93–102.

89. Ho CH, Eastman C, Catrambone R. An investigation of 2D and 3D spatial and mathematical abilities. Design Studies. 2006;27:505–524.
90. Ho HZ, Senturk D, Lam AG, Zimmer JM, Hong S, Okamoto Y, et al. The affective and cognitive dimensions of math anxiety: A cross-national study. Journal for Research in Mathematics Education.2000;31:362–379.
91. Hong E, Aqui Y. Cognitive and motivational characteristics of adolescents gifted in mathematics: Comparisons among students with different types of giftedness. Gifted Child Quarterly.2004;48:191–201.
92. Hosenfeld I, Koller O, Baumert J. Why sex differences in mathematics achievement disappear in German secondary schools: A reanalysis of the German TIMSS-data. Studies in Educational Evaluation. 1999;25:143–161.
93. Huang J. An investigation of gender differences in cognitive abilities among Chinese high school students. Personality and Individual Differences. 1993;15:717–719.
94. Iben MF. Attitudes and mathematics. Comparative Education. 1991;27:135–151.
95. Isiksal M, Askar P. The effect of spreadsheet and dynamic geometry software on the achievement and self-efficacy of 7th-grade students. Educational Research. 2005;47:333–350.
96. Jinabhai CC, Taylor M, Rangongo MF, Mkhize NJ, Anderson S, Pillay BJ, et al. Investigating the mental abilities of rural Zulu primary school children in South Africa. Ethnicity & Health.2004;9:17–36. [PubMed]
97. Jones ML, Rowsey RE. The effects of immediate achievement and retention of middle school students involved in a metric unit designed to promote the development of estimating skills. Journal of Research in Science Teaching. 1990;27:901–913.
98. Kahn M. A class act - mathematics as filter of equity in South Africa's schools. Perspectives in Education. 2005;23:139–148.
99. Kaiser J. The role of family configuration, income, and gender in the academic achievement of young self-care children. Early Child Development and Care. 1994;97:91–105.
100. Kass RG, Fish JM. Positive reframing and the test performance of test anxious children. Psychology in the Schools. 1991;28:43–52.
101. Kee DW, Gottfried A, Bathurst K. Consistency of hand preference: Predictions to intelligence and school achievement. Brain and Cognition. 1991;16:1–10. [PubMed]
102. Keller J. Blatant stereotype threat and women's math performance: Self-handicapping as a strategic means to cope with obtrusive negative performance expectations. Sex Roles. 2002;47:193–198.
103. Kelly-Vance L, Caster A, Ruane A. Non-graded versus graded elementary schools: An analysis of achievement and social skills. Alberta Journal of Educational Research. 2000;46:372–390.
104. Kenney-Benson GA, Pomerantz EM, Ryan AM, Patrick H. Sex differences in math performance: The role of children's approach to schoolwork. Developmental Psychology. 2006;42:11–26.[PubMed]
105. Kiger DM. The effect of group test-taking environment on standardized achievement test scores: A randomized block field trial. American Secondary Education. 2005;33:63–72.
106. Kimura D. Body asymmetry and intellectual pattern. Personality and Individual Differences.1994;17:53–60.
107. Kloosterman P. Beliefs and achievement in seventh-grade mathematics. Focus on Learning Problems in Mathematics. 1991;v13:3.
108. Koizumi R. The relationship between perceived attainment and optimism, and academic achievement and motivation. Japanese Psychological Research. 1992;34:1–9.

109. Kontrová J, Palkovicová E, Árochová O. Load and stress in the teaching process. Studia Psychologica. 1991;33:129–137.
110. Kumar S, Harizuka S. Cooperative learning-based approach and development of learning awareness and achievement in mathematics in elementary school. Psychological Reports. 1998;82:587–591.
111. Kwok DC, Lytton H. Perceptions of mathematics ability versus actual mathematics performance: Canadian and Hong Kong Chinese children. British Journal of Educational Psychology.1996;66:209–222. [PubMed]
112. Lachance JA, Mazzocco MMM. A longitudinal analysis of sex differences in math and spatial skills in primary school age children. Learning and Individual Differences. 2006;16:195–216.[PMC free article] [PubMed]
113. Lagace DC, Kutcher SP, Robertson HA. Mathematics deficits in adolescents with bipolar I disorder. American Journal of Psychiatry. 2003;160:100–104. [PubMed]
114. Lakes KD, Hoyt WI. Promoting self-regulation through school-based martial arts training. Journal of Applied Developmental Psychology. 2004;25:283–302.
115. Landgren M, Kjellman B, Gillberg C. “A school for all kinds of minds” - The impact of neuropsychiatric disorders, gender and ethnicity on school-related tasks administered to 9–10-year-old children. European Child & Adolescent Psychiatry. 2003;12:162–171. [PubMed]
116. Lau S, Leung K. Relations with parents and school and Chinese adolescents' self-concept, delinquency, and academic performance. British Journal of Educational Psychology. 1992;62:193–202. [PubMed]
117. LeFevre J, Kulak AG, Heymans SL. Factors influencing the selection of university majors varying in mathematical content. Canadian Journal of Behavioural Science. 1992;24:276–289.
118. Leonard J. How group composition influenced the achievement of sixth-grade mathematics students. Mathematical Thinking and Learning. 2001;3:175–200.
119. Lesko AC, Corpus JH. Discounting the difficult: How high math-identified women respond to stereotype threat. Sex Roles. 2006;54:113–125.
120. Lim TK. Gender-related differences in intelligence: Application of confirmatory factor analysis. Intelligence. 1994;19:179.
121. Lindblad F, Lindahl M, Theorell T, von Scheele B. Physiological stress reactions in 6th and 9th graders during test performance. Stress and Health. 2006;22:189–195.
122. Lindsay G, Desforges M. The use of the infant Index/Baseline-PLUS as a baseline assessment measure of literacy. Journal of Research in Reading. 1999;22:55–66.
123. Lloyd JEV, Walsh J, Yailagh MS. Sex differences in performance attributions, self-efficacy, and achievement in mathematics: If I'm so smart, why don't I know it? Canadian Journal of Education.2005;28:384–408.
124. Lopez CL, Sullivan HJ. Effect of personalization of instructional context on the achievement and attitudes of Hispanic students. Educational Technology Research and Development. 1992;40:5–13.
125. López CL, Sullivan HJ. Effects of personalized math instruction for Hispanic students. Contemporary Educational Psychology. 1991;16:95–100.
126. Lopez-Sobaler AM, Ortega RM, Quintas ME, Navia B, Requejo AM. Relationship between habitual breakfast and intellectual performance (logical reasoning) in well-nourished schoolchildren of Madrid (Spain) European Journal of Clinical Nutrition. 2003;57:49–53. [PubMed]
127. Low R, Over R. Gender differences in solution of algebraic word problems containing irrelevant information. Journal of Educational Psychology. 1993;85:331–339.

128. Lowrie T, Kay R. Relationship between visual and non-visual solution methods and difficulty in elementary mathematics. Journal of Educational Research. 2001;94:248–255.
129. Lubinski D, Humphreys LG. A broadly based analysis of mathematical giftedness. Intelligence.1990;14:327–355.
130. Lummis M, Stevenson HW. Gender differences in beliefs and achievement: A cross-cultural study. Developmental Psychology. 1990;26:254–263.
131. Lyons JB, Schneider TR. The influence of emotional intelligence on performance. Personality and Individual Differences. 2005;39:693–703.
132. Manger T, Eikeland OJ. Relationship between boys' and girls' nonverbal ability and mathematical achievement. School Psychology International. 1996;17:71–80.
133. Manger T, Eikeland OJ. The effect of mathematics self-concept on girls' and boys' mathematical achievement. School Psychology International. 1998;19:5–18.
134. Manger T, Eikeland OJ. The effects of spatial visualization and students' sex on mathematical achievement. British Journal of Psychology. 1998;89:17–25. [PubMed]
135. Manger T. Gender differences in mathematical achievement at the Norwegian elementary-school level. Scandinavian Journal of Educational Research. 1995;39:257–269.
136. Manger T, Eikeland O. Gender differences in mathematical sub-skills. Research in Education.1998;59:59–68.
137. Manger T, Gjestad R. Gender differences in mathematical achievement related to the ratio of girls to boys in school classes. International Review of Education. 1997;43:193.
138. Maqsud M. Effects of metacognitive skills and nonverbal ability on academic achievement of high school pupils. Educational Psychology. 1997;17:387–397.
139. Maqsud M, Khalique CM. Relationships of some socio-personal factors to mathematics achievement of secondary school and university students in Bophuthatswana. Educational Studies in Mathematics. 1991;22:377–390.
140. Maqsud M, Rouhani S. Relationships between socioeconomic status, locus of control, self-concept, and academic achievement of Batswana adolescents. Journal of Youth and Adolescence.1991;20:107–114. [PubMed]
141. Maree JG, Erasmus CP. Mathematics skills of tswana-speaking learners in the north west province of South Africa. International Journal of Adolescence and Youth. 2006;13:71–97.
142. Matthews DJ. Diversity in domains of development: Research findings and their implications for gifted identification and programming. Roeper Review. 1997;19:172.
143. Mboya MM. Self-concept of academic ability as a function of sex, age, and academic achievement among African adolescents. Perceptual and Motor Skills. 1998;87:155–161. [PubMed]
144. McCoy LP. Effect of demographic and personal variables on achievement in eighth-grade algebra. Journal of Educational Research. 2005;98:131–135.
145. McKenzie B, Bull R, Gray C. The effects of phonological and visual-spatial interference on children's arithmetical performance. Educational and Child Psychology. 2003;20:93–108.
146. McIntyre RB, Lord CG, Gresky DM, Ten Eyck LL, Jay Frye GD, Bond CFJ. A social impact trend in the effects of role models on alleviating women's mathematics stereotype threat. Current Research in Social Psychology. 2005;10
147. McNiece R, Jolliffe F. An investigation into regional differences in educational performance in the national child development study. Educational Research. 1998;40:17–30.
148. Medina M., Jr. Spanish achievement in a maintenance bilingual education program: Language proficiency, grade and gender comparisons. Bilingual Research Journal. 1993;17:57.
149. Miller CJ, Crouch JG. Gender difference in problem-solving-expectancy and problem context. Journal of Psychology. 1991;125:327–336.

150. Mills CJ, Ablard KE, Gustin WC. Academically talented students' achievement in a flexibly paced mathematics program. Journal for Research in Mathematics Education. 1994;25:495–511.
151. Mills CJ, Ablard KE, Stumpf H. Gender differences in academically talented young students' mathematical reasoning: Patterns across age and sub-skills. Journal of Educational Psychology.1993;85:340–346.
152. Mohsin M, Nath SR, Chowdhury AMR. Influence of socioeconomic factors on basic competencies of children in Bangladesh. Journal of Biosocial Science. 1996;28:15–24. [PubMed]
153. Moodaley RR, Grobler AA, Lens W. Study orientation and causal attribution in mathematics achievement. South African Journal of Psychology. 2006;36:634–655.
154. Morrison FJ, Griffith EM, Alberts DM. Nature-nurture in the classroom: Entrance age, school readiness, and learning in children. Developmental Psychology. 1997;33:254–262. [PubMed]
155. Murphy LO, Ross SM. Protagonist gender as a design variable in adapting mathematics story problems to learner interests. Educational Technology Research and Development. 1990;38:27–37.
156. Mwamwenda TS. Sex differences in mathematics performance among african university students. Psychological Reports. 2002;90:1101–1104. [PubMed]
157. Narciss S, Huth K. Fostering achievement and motivation with bug-related tutoring feedback in a computer-based training for written subtraction. Learning and Instruction. 2006;16:310–322.
158. Nasser F, Birenbaum M. Modeling mathematics achievement of Jewish and Arab eighth graders in Israel: The effects of learner-related variables. Educational Research and Evaluation. 2005;11:277–302.
159. Nelson JR, Benner GJ, Lane K, Smith BW. Academic achievement of K-12 students with emotional and behavioral disorders. Exceptional Children. 2004;71:59–73.
160. Nyangeni NP, Glencross MJ. Sex differences in mathematics achievement and attitude toward mathematics. Psychological Reports. 1997;80:603–608.
161. Olszewski-Kubilius P, Turner D. Gender differences among elementary school-aged gifted students in achievement. Journal for the Education of the Gifted. 2002;25:233–268.
162. Onatsu-Arvilommi T, Nurmi JE. The role of task-avoidant and task-focused behaviors in the development of reading and mathematical skills during the first school year: A cross-lagged longitudinal study. Journal of Educational Psychology. 2000;92:478–491.
163. O'Neil HF, Abedi J, Miyoshi J. Monetary incentives for low-stakes tests. Educational Assessment.2005;10:185–208.
164. Ong W, Allison J, Haladyna TM. Student achievement of 3rd-graders in comparable single-age and multiage classrooms. Journal of Research in Childhood Education. 2000;14(2):205–215.
165. Opyene-Eluk P, Opolot-Okurut C. Gender and school-type differences in mathematics achievement of senior three pupils in central Uganda: An exploratory study. International Journal of Mathematical Education in Science and Technology. 1995;26:871–886.
166. Pajares F, Miller MD. Role of self-efficacy and self-concept beliefs in mathematical problem-solving – A path-analysis. Journal of Educational Psychology. 1994;86:193–203.
167. Pajares F. Mathematics self-efficacy and mathematical problem solving: Implications of using different forms of assessment. Journal of Experimental Education. 1997;65:213–228.
168. Pajares F. Self-efficacy beliefs and mathematical problem-solving of gifted students. Contemporary Educational Psychology. 1996;21:325–344. [PubMed]
169. Panchon A. The effects of white-noise and gender on mental task. Psychologia. 1994;37:234–240.
170. Park HS. Computational mathematical abilities of African American girls. Journal of Black Studies.1999;30:204–215.

171. Park H, Bauer SC. Gender differences among top performing elementary school students in mathematical ability. Journal of Research and Development in Education. 1998;31:133–141.
172. Pascarella ET, Bohr L, Nora A. Intercollegiate athletic participation and freshman-year cognitive outcomes. Journal of Higher Education. 1995;66:369–387.
173. Pehkonen E. Learning results from the viewpoint of equity: Boys, girls and mathematics. Teaching Mathematics and its Applications. 1997;16:58.
174. Pigg AE, Waliczek TM. Effects of a gardening program on the academic progress of third, fourth, and fifth grade math and science students. Horttechnology. 2006;16:262–264.
175. Pomplun M. Gender differences for constructed-response mathematics items. Educational and Psychological Measurement. 1999;59:597–614.
176. Pope GA, Wentzel C, Braden B. Relationships between gender and Alberta Achievement Test Scores during a four-year period. Alberta Journal of Educational Research. 2006;52:4–15.
177. Quinn DM. The interference of stereotype threat with women's generation of mathematical problem-solving strategies. Journal of Social Issues. 2001;57:55–71.
178. Rammstedt B. Self-estimated intelligence - gender differences, relationship to psychometric intelligence and moderating effects of level of education. European Psychologist. 2002;7:275–284.
179. Randhawa BS. Validity of performance assessment in mathematics for early adolescents. Canadian Journal of Behavioural Science-Revue Canadienne Des Sciences Du Comportement. 2001;33:14–24.
180. Randhawa BS. Self-efficacy in mathematics, attitudes, and achievement of boys and girls from restricted samples in two countries. Perceptual and Motor Skills. 1994;79:1011–1018.
181. Randhawa BS. Understanding sex differences in the components of mathematics achievement. Psychological Reports. 1993;73:435–444.
182. Reynolds AJ, Mehana M. Does preschool intervention affect childrens perceived competence. Journal of Applied Developmental Psychology. 1995;16:211–230.
183. Reys RE, Reys B. Computational estimation performance and strategies used by fifth- and eighth-grade Japanese students. Journal for Research in Mathematics Education. 1991;22:39–58.
184. Reys RE, Reys B. Mental computation performance and strategy use of japanese students in grades 2, 4, 6, and 8. Journal for Research in Mathematics Education. 1995;26:304–326.
185. Robinson NM, Abbott RD. The structure of abilities in math-precocious young children: Gender similarities and differences. Journal of Educational Psychology. 1996;88:341–352.
186. Rosselli M, Ardila A, Bateman JR. Neuropsychological test scores, academic performance., and developmental disorders in Spanish-speaking children. Developmental Neuropsychology.2001;20:355–373. [PubMed]
187. Rouxel G. Cognitive-affective determinants of performance in mathematics and verbal domains - gender differences. Learning and Individual Differences. 2001;12:287–310.
188. Rouxel G. Cognitive-affective determinants of performance in mathematics and verbal domains gender differences. Learning and Individual Differences. 2000;12:287–310.
189. Rudnitsky A, Etheredge S, Freeman SJM. Learning to solve addition and subtraction word problems through a structure-plus-writing approach. Journal for Research in Mathematics Education.1995;26:467–486.
190. Ruthven K. The influence of graphic calculator use on translation from graphic to symbolic forms. Educational Studies in Mathematics. 1990;21:431.
191. Saigal S, Lambert M, Russ C. Self-esteem of adolescents who were born prematurely. Pediatrics.2002;109:429–433. [PubMed]
192. Salawu AA. Relationship between adolescents' perception of parents' behaviour and their academic achievement. IFE Psychologia: An International Journal. 1993;1:153–165.

193. Salerno CA. The effect of time on computer-assisted instruction for at-risk students. Journal of Research on Computing in Education. 1995;28:85–97.
194. Sappington J, Larsen C, Martin J. Sex-differences in math problem-solving as a function of gender-specific item content. Educational and Psychological Measurement. 1991;51:1041–1048.
195. Sarouphim KM. Discover in middle school: Identifying gifted minority students. Journal of Secondary Gifted Education. 2004;15:61–69.
196. Schellinger T. Correlations among special-educations students WISC - RIQS and SRA scores. Perceptual and Motor Skills. 1991;73:1225–1226. [PubMed]
197. Seegers G. Gender-related differences in self-referenced cognitions in relation to mathematics. Journal for Research in Mathematics Education. 1996;27:215–240.
198. Sekaquaptewa D. Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. Journal of Experimental Social Psychology. 2003;39:68–74.
199. Shannon HD, Allen TW. The effectiveness of a REBT training program in increasing the performance of high school students in mathematics. Journal of Rational-Emotive & Cognitive Behavior Therapy. 1998;16:197–209.
200. Sheehan KR, Gray MW. Sex bias in the SAT and the DTMS. Journal of General Psychology.1992;119:5–14.
201. Shibley IA, Jr., Milakofsky L. College chemistry and Piaget: An analysis of gender difference, cognitive abilities, and achievement measures seventeen years apart. Journal of Chemical Education.2003;80:569–573.
202. Shymansky JA, Yore LD, Anderson JO. Impact of a school district's science reform effort on the achievement and attitudes of third- and fourth-grade students. Journal of Research in Science Teaching. 2004;41:771–790.
203. Skaalvik EM, Rankin RJ. Math, verbal, and general academic self-concept: The internal/external frame of reference model and gender differences in self-concept structure. Journal of Educational Psychology. 1990;82:546–554.
204. Skaalvik EM, Rankin RJ. Gender differences in mathematics and verbal achievement, self-perception and motivation. British Journal of Educational Psychology. 1994;64:419–428.
205. Skaggs G, Lissitz RW. The consistency of detecting item bias across different test administrations: Implications of another failure. Journal of Educational Measurement. 1992;29:227–242.
206. Slate JR, Jones CH, Turnbough R, Bauschlicher L. Gender differences in achievement scores on the metropolitan achievement test-6 and the stanford achievement test-8. Research in the Schools.1994;1:59–62.
207. Smees R, Sammons P, Thomas S, Mortimore P. Examining the effect of pupil background on primary and secondary pupils' attainment: Key findings from the improving school effectiveness project. Scottish Educational Review. 2002;34:6.
208. Smith JL, White PH. Development of the domain identification measure: A tool for investigating stereotype threat effects. Educational and Psychological Measurement. 2001;61:1040–1057.
209. Spencer SJ, Steele CM, Quinn DM. Stereotype threat and women's math performance. Journal of Experimental Social Psychology. 1999;35:4–28.
210. Srivastava NC. Verbal test of intelligence as a predictor of success in science and mathematics. Psycho-Lingua. 1993;23:65–70.
211. Stage FK, Kloosterman P. Gender, beliefs, and achievement in remedial college-level mathematics. Journal of Higher Education. 1995;66:294–311.
212. Standing LG, Sproule RA, Leung A. Can business and economics students perform elementary arithmetic? Psychological Reports. 2006;98:549–555. [PubMed]

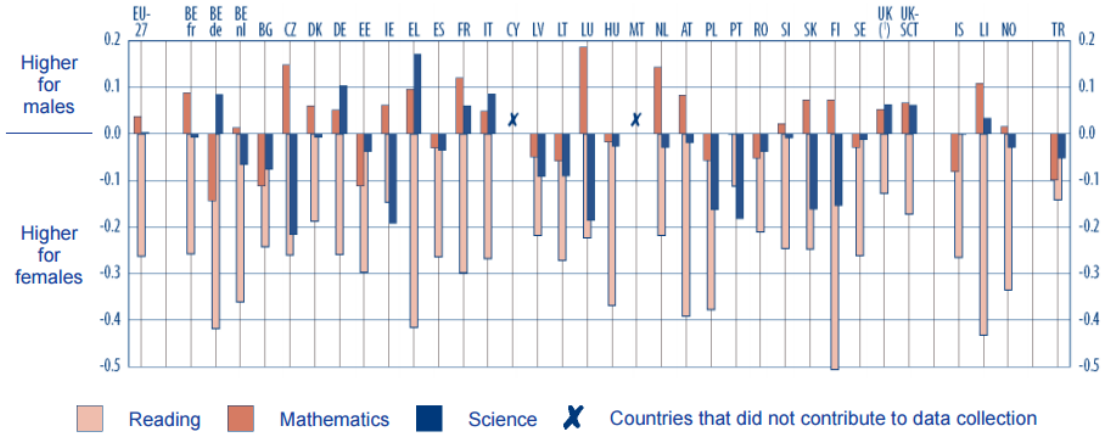
213. Stevenson HW, Chen C, Booth J. Influences of schooling and urbanural residence on gender differences in cognitive abilities and academic achievement. Sex Roles. 1990;23:535–551.
214. Stricker LJ, Ward WC. Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. Journal of Applied Social Psychology. 2004;34:665–693.
215. Stumpf H, Haldimann M. Spatial ability and academic success of sixth grade students at international schools. School Psychology International. 1997;18:245–259.
216. Subotnik RF, Strauss SM. Gender differences in classroom participation and achievement: An experiment involving advanced placement calculus classes. Journal of Secondary Gifted Education.1995;6:77.
217. Swiatek MA, Lupkowski-Shoplik A, O'Donoghue CC. Gender differences in above-level EXPLORE scores of gifted third through sixth graders. Journal of Educational Psychology. 2000;92:718–723.
218. Tartre LA, Fennema E. Mathematics achievement and gender: A longitudinal study of selected cognitive and affective variables {grades 6–12} Educational Studies in Mathematics. 1995;28:199–217.
219. Taylor L. An integrated learning system and its effect on examination performance in mathematics. Computers & Education. 1999;32:95–107.
220. Thompson GW, et al. Gender differences in an experimental program on arithmetic problem solving and computation. Midwestern Educational Researcher. 1992;5:20.
221. Tiedemann J, Faber G. Preschoolers' maternal support and cognitive competencies as predictors of elementary achievement. Journal of Educational Research. 1992;85:348–354.
222. Travis B, Lennon E. Spatial skills and computer-enhanced instruction in calculus. The Journal of Computers in Mathematics and Science Teaching. 1997;16:467–475.
223. Tsui M, Rich L. The only child and educational opportunity for girls in urban china. Gender & Society. 2002;16:74–92.
224. Undheim JO, Nordvik H, Gustafsson K, Undheim AM. Academic achievements of high-ability students in egalitarian education: A study of able 16-year-old students in Norway. Scandinavian Journal of Educational Research. 1995;39:157–167.
225. Valanides NC. Formal reasoning and science teaching. School Science and Mathematics.1996;96:99–107.
226. VanDerHeyden AM, Broussard C, Cooley A. Further development of measures of early math performance for preschoolers. Journal of School Psychology. 2006;44:533–553.
227. Vermeer HJ, Boekaerts M, Seegers G. Motivational and gender differences: Sixth-grade students' mathematical problem-solving behavior. Journal of Educational Psychology. 2000;92:308–315.
228. Walsh M, Hickey C, Duffy J. Influence of item content and stereotype situation on gender differences in mathematical problem solving. Sex Roles. 1999;41:219–240.
229. Wang N, Lane S. Detection of gender-related differential item functioning in a mathematics performance assessment. Applied Measurement in Education. 1996;9:175–199.
230. Wangu RS, Thomas KJ. Attitude towards and achievement in mathematics among high school students of tribal town of Aizawl. Indian Journal of Psychometry & Education. 1995;26:31–36.
231. Warwick DP, Jatoi H. Teacher gender and student achievement in Pakistan. Comparative Education Review. 1994;38:377–399.
232. Watt HMG. Measuring attitudinal change in mathematics and english over the 1st year of junior high school: A multidimensional analysis. Journal of Experimental Education. 2000;68:331–361.
233. Watt HMG. The role of motivation in gendered educational and occupational trajectories related to maths. Educational Research and Evaluation. 2006;12:305–322.

234. Watt HMG, Bornholt LJ. Social categories and student perceptions in high school mathematics. Journal of Applied Social Psychology. 2000;30:1492–1503.
235. Werdelin I. Sex differences in performance scores and patterns of development. Interdisciplinaria Revista De Psicología y Ciencias Afines. 1996;13:35–65.
236. Williams JE, Montgomery D. Using frame of reference theory to understand the self-concept of academically able students. Journal for the Education of the Gifted. 1995;18:400–409.
237. Witt EA, Dunbar SB, Hoover HD. A multivariate perspective on sex differences in achievement and later performance among adolescents. Applied Measurement in Education. 1994;7:241–254.
238. Wu M, Greenan JP. The effects of a generalizable mathematics skills instructional intervention on the mathematics achievement of learners in secondary CTE programs. Journal of Industrial Teacher Education. 2003;40:23–50.
239. Xu J, Farrell EW. Mathematics performance of shanghai high school students: A preliminary look at gender differences in another culture. School Science and Mathematics. 1992;92:442–445.
240. Yadrick RM, Regian JW, RobertsonSchule L, Gomez GC. Interface, instructional approach, and domain learning with a mathematics problem-solving environment. Computers in Human Behavior.1996;12:527–548.
241. Zervas Y. Effect of a physical exercise session on verbal, visuospatial, and numerical ability. Perceptual and Motor Skills. 1990;71:379–383

8 Appendices

8.1 Appendix - Relevance of subject perception, PISA 2006 results

Figure 2.1: Gender difference (M-F) in perceived importance of doing well in reading, mathematics and science for 15 year-old pupils, 2006



	EU-27	BE fr	BE de	BE nl	BG	CZ	DK	DE	EE	IE	EL	ES	FR	IT	CY	LV	LT	LU
	-0.26	**-0.26**	**-0.42**	**-0.36**	**-0.24**	**-0.26**	**-0.19**	**-0.26**	**-0.30**	**-0.15**	**-0.42**	**-0.26**	**-0.30**	**-0.27**	✗	**-0.22**	**-0.27**	**-0.22**
	0.04	**0.09**	**-0.14**	0.01	**-0.11**	**0.15**	**0.06**	**0.05**	**-0.11**	**0.06**	**0.10**	-0.03	**0.12**	**0.05**	✗	**-0.05**	**-0.06**	**0.19**
	0.00	-0.01	0.08	**-0.07**	**-0.08**	**-0.22**	-0.01	**0.10**	-0.04	**-0.19**	**0.17**	-0.04	0.06	**0.09**	✗	**-0.09**	**-0.09**	**-0.19**

	HU	MT	NL	AT	PL	PT	RO	SI	SK	FI	SE	UK (1)	UK-SCT		IS	LI	NO	TR
	-0.37	✗	**-0.22**	**-0.39**	**-0.38**	**-0.11**	**-0.21**	**-0.25**	**-0.25**	**-0.51**	**-0.26**	**-0.13**	**-0.17**		**-0.27**	**-0.43**	**-0.34**	**-0.14**
	-0.02	✗	**0.14**	**0.08**	**-0.06**	0.00	-0.05	0.02	**0.07**	**0.07**	-0.03	**0.05**	0.07		**-0.08**	0.11	0.02	**-0.10**
	-0.03	✗	-0.03	-0.02	**-0.16**	**-0.18**	-0.04	-0.01	**-0.16**	**-0.15**	-0.01	**0.06**	0.06		0.00	0.03	-0.03	-0.05

UK (1) = UK-ENG/WLS/NIR.

Source: OECD, PISA 2006 database.

Explanatory notes

The results are based on answers to the question: 'in general, how important do you think it is for you to do well in the subject below?' with four answer categories: very important, important, of little importance and not important at all. The graph shows coefficients of three different simple linear regression models.

For further information on the PISA survey, see the Glossary.

Values that are statistically significant (p<.05) are indicated in **bold**.

8.2 Appendix – Additional summary statistics

8.2.1 Other control variable statistics

	Obs	Description
Timed	*294*	49 untimed (17%), 245 (83%) timed
Format	*230*	86% was pencil & paper, 5% computerized and 8% oral/behavioural
Type	*189*	50% includes multiple choice, 57% includes short answer and 19% includes open response items.
Content	*205*	78% includes number and operations items, 26% include algebra items, 44% include geometry items, 18% include measurement items and 14% include data analysis & probability items.
Depth of Knowledge	*120*	74% includes items with depth of knowledge level 1, 43% includes items with depth of knowledge level 2 and7,6% includes items with depth of knowledge level 3. Of concern is that none of the 120 include includes items with depth of knowledge level 4
Stakes	*272*	133 studies of these (49%) are considered low-stakes test, 54 (20%) studies used high-stake tests to retrieve data and 85 are undefined.
National	*441*	U.S. 229 Canada 24 Mexico/Caribbean & Central/South America 9 Europe 75 Australia/New Zealand 15 Asia 45 Unreported/Mixed 1 Africa 25 Middle East 18

8.2.2 All statistics

Variable	Obs	Mean	Std. Dev.	Min	Max	Description
es	435	0.10	0.38	-2.17	1.79	Unbiased effect size
vr	369	1.15	0.48	0.00	5.96	Raw variance ratio
year	435	1997.24	5.21	1990	2006	Year of publication
source	435	1.20	0.49	1.00	3.00	Source of test code: article description only, article description & examples or actual items
n_male_A	422	1328.54	14148.94	1.00	276041.00	Number of males
m_male_A	402	98.06	515.54	-0.17	9933.00	Mean score for males
sd_male_A	372	15.22	77.16	0.00	1480.00	Standard deviation for males
n_female_A	422	794.94	5594.51	5.00	93637.00	Number of females
m_female_A	402	97.36	519.22	-0.44	10008.00	Mean score for females
sd_femal_A	371	14.48	70.36	0.06	1345.00	Standard deviation for females
n_A	435	2937.01	18809.83	12.00	320816.00	Total n (female + male)
sdp	370	14.98	73.93	0.06	1413.70	
n_male_L	422	1329.06	14148.92	1.00	276041.00	Number of males
m_male_L	397	99.27	518.68	-0.17	9933.00	Mean score for males
sd_male_L	369	15.33	77.47	0.00	1480.00	Standard deviation for males
n_female_L	422	794.97	5594.51	5.00	93637.00	Number of females
m_female_L	397	98.59	522.37	-0.44	10008.00	Mean score for females

sd_femal_L	369	14.58	70.54	0.06	1345.00	Standard deviation for females
n_L	435	2937.93	18809.74	12.00	320816.00	Total n (female + male)
curric	430	0.30	0.46	0.00	1.00	Curricular focus of test (0=standard assessment, 1=classroom, 9=uncodable)
format1	228	0.86	0.34	0.00	1.00	Paper & pencil
format2	227	0.06	0.23	0.00	1.00	Computerized
format3	227	0.08	0.27	0.00	1.00	Oral/Behavioural
type1	187	0.50	0.50	0.00	1.00	Includes multiple choice items
type2	187	0.57	0.51	0.00	2.00	Includes short answer items
type3	186	0.19	0.39	0.00	1.00	Includes open response items
content1	202	0.78	0.42	0.00	1.00	Includes number & operations items
content2	202	0.27	0.44	0.00	1.00	Includes algebra items
content3	201	0.44	0.50	0.00	1.00	Includes geometry items
content4	201	0.19	0.39	0.00	1.00	Includes measurement items
content5	201	0.13	0.34	0.00	1.00	Includes data analysis & probability items
dok1	119	0.74	0.44	0.00	1.00	Includes depth of knowledge = 1 items
dok2	119	0.44	0.50	0.00	1.00	Includes depth of knowledge = 2 items
dok3	118	0.08	0.27	0.00	1.00	Includes depth of knowledge = 3 items
dok4	118	0.00	0.00	0.00	0.00	Includes depth of knowledge = 4 items
dok_high	434	6.99	3.32	1.00	9.00	Highest depth of knowledge level of any test item
stakes_A	270	3.00	4.06	0.00	9.00	High or low stakes (1= high, 0=low, 9=unknown)
timed_A	292	0.84	0.37	0.00	1.00	Was the test timed? Dummy.
timed_L	136	0.79	0.43	0.00	2.00	Was the test timed? Dummy.
minmax	179	45.88	40.51	1.50	195.00	Maximum amount of minutes allowed on test
noq	298	41.08	31.75	3.00	240.00	Number of questions on test
da	435	0.10	0.38	-2.26	1.77	Updated raw effect size
d	435	0.10	0.39	-2.30	1.79	Raw effect size
age	430	3.46	1.07	1.00	6.00	Age group of sample
se	422	0.19	0.13	0.00	1.03	Standard error of the effect size
w	422	423.17	3017.65	0.94	41933.72	Inverse variance weight
ltvr	368	0.07	0.39	-2.12	1.78	Variance ratio (log transformed for weighted average)
sevr	368	0.19	0.13	0.01	0.68	
wvr_A	368	341.46	2393.44	2.15	38494.65	

8.3 Appendix – Visual insight into missing variable ‘ depth of knowledge’

Figure 28 and Figure 29 are a visualised example of how the groups with and without observations of missing control variables (in this case the control variable ‘depth of knowledge) have significantly different test lengths.

Figure 28: histograms of studies with missing values for the depth of knowledge variable (group 0) versus those with a value for this variable (group 1). The maximum amount of minutes of all observations in the two groups is visualized.

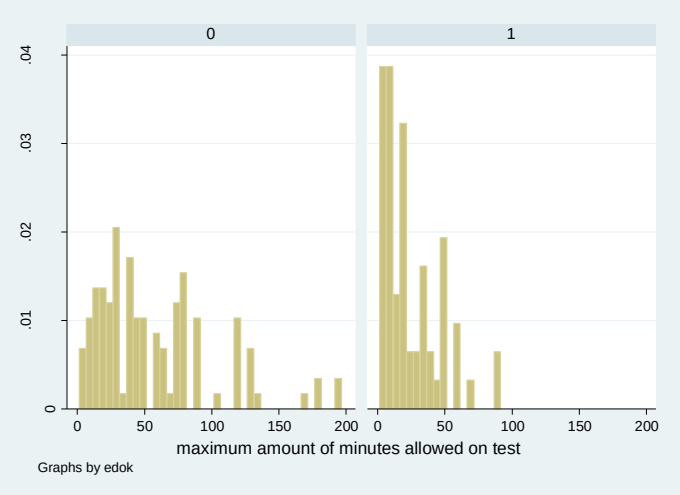
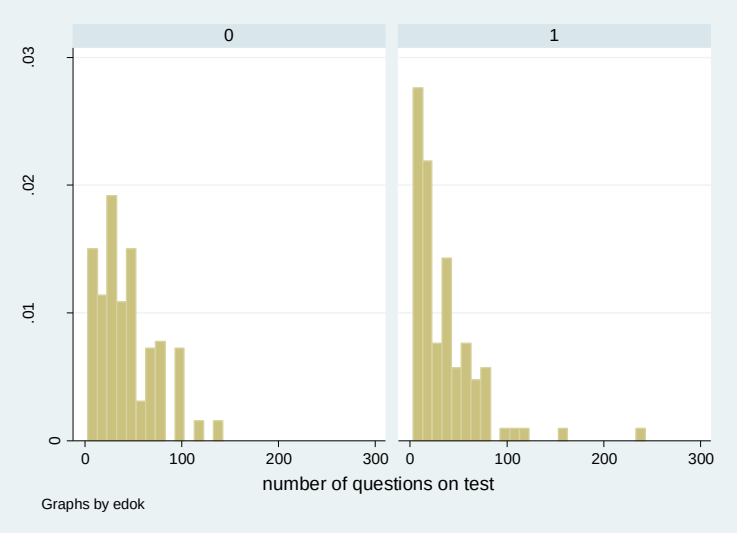


Figure 29: histograms of studies with missing values for the depth of knowledge variable (group 0) versus those with a value for this variable (group 1). The number of questions of all observations in the two groups is visualized.



8.4 Appendix - Visual representation weight article 87

Figure 30 and Figure 31 illustrate the impact of excluding the heavy weighting navy article that suffers from selection bias from the regressions of the maximum amount of minutes allowed on a test on the standardized gender difference (see section 3.2). Figure 32 and Figure 33 do the same for a regression with the number of questions as the independent variable of interest.

Figure 30: Weighted scatterplot of standardized difference on the maximum amount of minutes, including article 87.

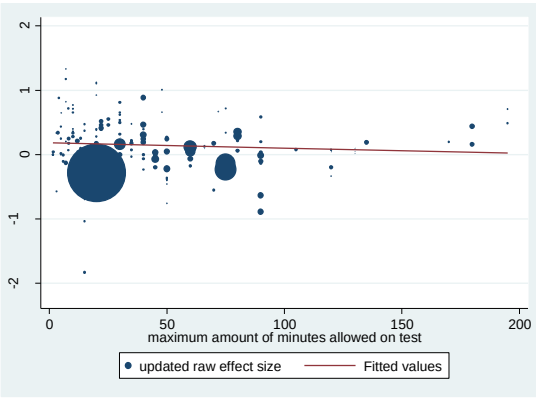


Figure 31: Weighted scatterplot of standardized difference on the maximum amount of minutes, excluding article 87

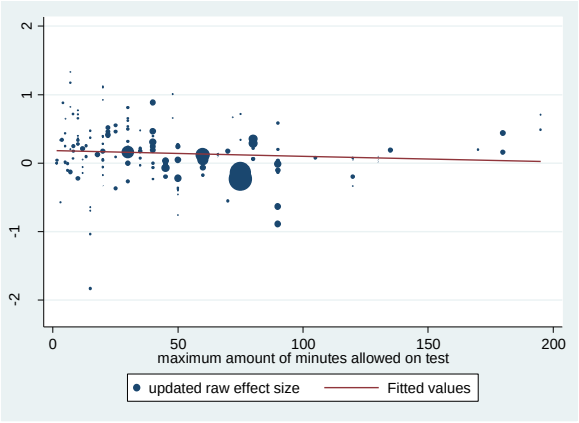


Figure 32: Weighted scatterplot of standardized difference on the number of questions, including article 87.

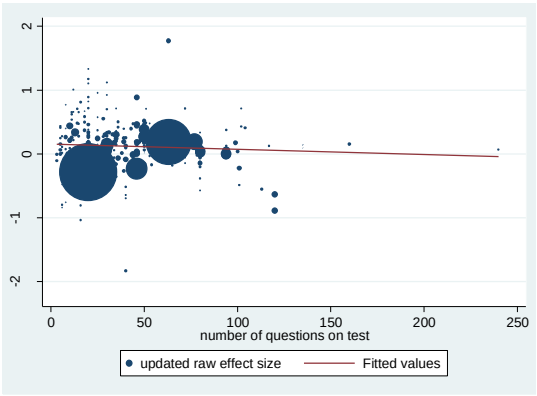
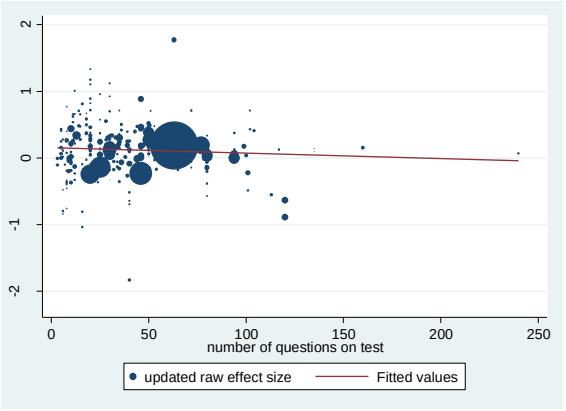


Figure 33: Weighted scatterplot of standardized difference on the number of questions, excluding article 87.



8.5 Appendix - Summary findings as provided by Lindberg et al. (2010)

Table 2

Summary of Meta-Analysis Results (Study 1), as Moderated by Sample Characteristics

Sample characteristic	*d*	*k*	*p*
Ability			***
Low ability	+0.07	15	
General ability	+0.07	304	
Moderately selective	+0.15	79	
Highly selective	+0.40	27	
Nationality			
U.S.	+0.10	226	
Canada	+0.01	13	
Central/South America & Mexico	−0.06	11	
Europe	+0.07	70	
Australia & New Zealand	+0.10	15	
Asia	+0.17	45	
Africa	+0.21	26	
Middle East	+0.12	16	
Ethnicity (U.S. samples)			**
Primarily European American	+0.13	58	
Primarily minority (combined)[a]	−0.05	32	
Age			***
Preschool	−0.15	3	
Elementary school	+0.06	86	
Middle school	−0.00	140	
High school	+0.23	110	
College	+0.18	78	
Adult	−0.07	7	

Note. *k* = number of studies; *d* = effect size. Studies that provided insufficient information to code a particular moderator are omitted from that analysis. Therefore, *k* fluctuates between analyses, and the results of moderator analyses do not represent the full body of studies used to compute the overall mean effect size reported in Study 1.

[a] The number of samples for distinct U.S. ethnic groups was too small to analyze effects for each group separately, so all ethnic minority samples were combined.

** $p < .01$. *** $p < .001$.

Table 1

Weighted Ordinary Least Squares Regressions Predicting Gender Differences in Math Performance, as Moderated by Test Characteristics (Study 1)

Moderator (*k* affirmative)	β	*k*	R^2	*p*
Problem type		189	.04	*
Contains multiple choice (95)	+.16			
Contains short answer (105)	−.03			
Contains open-ended (36)	−.09			
Problem content		205	.04	
Contains numbers & operations (160)	+.02			
Contains algebra (54)	+.10			
Contains geometry (90)	+.14			
Contains measurement (38)	−.08			
Contains data analysis & probability (29)	+.09			
Problem depth of knowledge		120	.01	
Contains Level 1 (89)	−.10			
Contains Level 2 (52)	+.00			
Contains Level 3 or 4 (9)	−.10			
Test time limit		137	.02	
Yes (105)	+.12			
Test curriculum focused		423	.01	
Yes (132)	−.08			

Note. *k* = number of studies; *d* = effect size. Studies that provided insufficient information to code a particular moderator are omitted from that analysis. Therefore, *k* fluctuates between analyses, and the results of moderator analyses do not represent the full body of studies used to compute the overall mean effect size reported in Study 1. Positive betas denote increases in effect size (advantage for males) as the value of the predictor increases, whereas negative betas denote decreases in effect size (advantage for females) as the value of the predictor increases.

* $p < .05$.

Table 3

Mean Weighted Effect Sizes, by Highest Depth of Knowledge Tested and Age Group

Depth of knowledge	Elementary	Middle school	High school	College	Adult
Level 1	+0.03 (*k* = 21)	−0.02 (*k* = 16)	+0.21 (*k* = 9)	+0.24 (*k* = 11)	+0.17 (*k* = 3)
Level 2	+0.12 (*k* = 5)	−0.07 (*k* = 16)	+0.37 (*k* = 16)	+0.12 (*k* = 8)	−0.28 (*k* = 1)
Level 3 or 4		−0.10 (*k* = 1)	+0.16 (*k* = 3)	−0.11 (*k* = 5)	

8.6 Appendix – Expanding the regression per variable

Observations are unweighted and clustered standard errors based on article number are used.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Effect on standardized gender gap											
Maximum allowed minutes	-.0008 (.0010)	-.0007 (.0010)	-.0010 (.0011)	-.0010 (.0010)	-.0013 (.0009)	-.0006 (.0012)	-.0002 (.0012)	-.0010 (.0012)	-.0009 (.0011)	-.0004 (.0013)	-.0005 (.0013)
Number of observations	179	179	179	179	179	179	179	179	179	179	179
Timed test	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year trend	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ability	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nationality	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Stakes	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes
Answer type	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes
Test format	No	No	No	No	No	No	No	No	Yes	Yes	Yes
Content type	No	No	No	No	No	No	No	No	No	Yes	Yes
Depth of Knowledge	No	No	No	No	No	No	No	No	No	No	Yes
Effect on standardized gender gap											
Number of questions	-.0008 (.0009)	-.0008 (.0009)	-.0010 (.0008)	-0009 (.0007)	-.0007 (.0007)	-.0002 (.0008)	-.0002 (.0007)	-.0003 (.0007)	-.0004 (.0007)	-.0004 (.0007)	-.0004 (.0007)
Number of observations	298	298	298	298	298	298	298	298	298	298	298
Timed test	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year trend	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ability	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nationality	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Stakes	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes
Answer type	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes
Test format	No	No	No	No	No	No	No	No	Yes	Yes	Yes
Content type	No	No	No	No	No	No	No	No	No	Yes	Yes
Depth of Knowledge	No	No	No	No	No	No	No	No	No	No	Yes