Sex Roles (2021) 85:481–496
https://doi.org/10.1007/s11199-021-01234-5

ORIGINAL ARTICLE

The Gender Similarities Hypothesis: Insights From A Multilevel Analysis of High-Stakes Examination Results in Mathematics

Ian Cantley[1] · James McAllister[1]

Accepted: 30 June 2021 / Published online: 12 July 2021

Abstract

The current study involved multilevel analysis of high-stakes examination results (i.e., GCSE) in Northern Ireland to investigate gender differentials in mathematical achievement, whereas most previous research in the area used results from low-stakes tests (i.e., PISA, TIMSS). The analysis supported the gender similarities hypothesis with respect to both overall and content domain-specific mathematical attainment. Similar conclusions were drawn from the current study as have been reported in studies into gender differentials using data from low-stakes assessments in the respective jurisdiction. This suggests that previously expressed concerns in the literature about the viability of using data derived from low-stakes assessments to accurately assess gender differentials in achievement may be unfounded. Furthermore, the context for the current study permitted an investigation into the effects of school type (grammar versus non-grammar) and gender on overall and domain-specific mathematical achievement, an area that has received scant attention in the literature. School type was not found to have an effect on the applicability of the gender similarities hypothesis with respect to mathematical achievement. The study findings are likely to prove useful to researchers and policymakers who are interested in gender equity issues in mathematics.

Keyword Gender equity · Mathematics · High-stakes assessments · Domain-specific achievement

Mathematical proficiency is considered an important skill to acquire as part of a good education. This is because of its pivotal role in everyday life and in many vocations (Leder & Forgasz, 2018; Valero, 2017). The issue of gender differences in mathematical performance is an important and contested one, particularly against the backdrop of the stereotypical view that women lack mathematical ability (Franceschini et al., 2014).

The analysis reported upon in this paper is used to test the gender similarities hypothesis (Hyde, 2005, 2014, 2016) in relation to student achievement in high-stakes mathematics examinations in Northern Ireland. Based on a review of 46 meta-analyses, Hyde (2005) concluded that, "males and females are similar on most, but not all, psychological variables" (p. 581). In her work, Hyde (2005) applied Cohen's d standardised measure of effect size to assess gender differences in psychological variables, with positive values of d favouring boys and men and negative values favouring girls and women. The following ranges were used to evaluate the magnitudes of effects: $|d| \leq .10$ is close-to-zero; $.10 < |d| \leq .35$ is small; $.35 < |d| \leq .65$ is moderate; $.65 < |d| \leq 1.00$ is large; and $|d| > 1.00$ is very large. Hyde's (2005) research revealed close-to-zero or small gender differences in mathematics-related constructs, with the exception of spatial ability, which had moderate to large effect sizes in favour of boys. The small, or close-to-zero, gender differences in mathematical performance have been supported by more recent meta-analytic studies (e.g., Lindberg et al., 2010), but national gender differentials have been shown to exhibit substantial variability (Else-Quest et al., 2010; Nollenberger et al., 2016; Rodríguez-Planas & Nollenberger, 2018). It is important to note that the research described here was almost exclusively conducted in the United States, either as a nation or in specific states or school districts. By contrast, the use of Northern Ireland as a site for the current study offers a distinct context to investigate gender differentials in mathematical achievement.

A large proportion of the recent research on gender differentials in mathematical achievement has been based on

✉ Ian Cantley
i.cantley@qub.ac.uk

James McAllister
jmcallister22@qub.ac.uk

[1] School of Social Sciences, Education and Social Work, Queen's University Belfast, 69-71 University Street, Belfast BT7 1HL, UK

analyses of student performance in large-scale international assessments such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS) (e.g., Ayalon & Livneh, 2013; Brunner et al., 2013; Demir et al., 2010; Else-Quest et al., 2010; Innabi & Dodeen, 2018; Liu & Wilson, 2009; Louis & Mistele, 2012; Meggiolaro, 2018). Given that international large-scale assessments have no direct consequences for participants, it has been suggested that their low-stakes nature can negatively impinge upon the amount of effort that students invest in their completion. This may compromise the validity and reliability of the associated ability measurements (Eklöf, 2010), which has the potential to negatively impact the reliability of gender differentials in achievement derived from large-scale international assessments.

Gender differentials in mathematical achievement have been shown to vary according to mathematical content domains such as number work, algebra, geometry and measures, and statistics and probability (e.g., Li et al., 2018; McGraw et al., 2006; Taylor & Lee, 2012). For example, a large male advantage in geometry has been reported by a number of scholars (e.g., Leahey & Guo, 2001; McGraw et al., 2006; Taylor & Lee, 2012). However, less attention has been paid to this issue than overall mathematical performance. The current article addresses this gap in the literature. This is an important area for investigation because some existing gender-focused research has demonstrated that women's study of, and achievement in, different content domains influence their uptake of the physical sciences, engineering, mathematics, and computer science in tertiary level education (Nix & Perez-Felkner, 2019; Perez-Felkner et al., 2012; Watt et al., 2012).

Background and Study Context

To further illuminate the gaps in existing work, this article reports on an analysis of a sample of high-stakes public examination results in mathematics from General Certificate of Secondary Education (GCSE) mathematics examinations, which are taken at the end of compulsory post-primary education in Northern Ireland, and have a pivotal role in determining students' future educational and vocational pathways. The analysis has been undertaken for both lower-ability and higher-ability students, who sit GCSE mathematics at foundation tier and higher tier respectively. Foundation tier examinations tend to be taken by lower-ability students who do not pursue STEM-related courses after GCSE, while higher tier examinations are taken by higher-ability students who may pursue STEM-related courses beyond GCSE. The analysis of gender differences in mathematical achievement by content domain (number, algebra, geometry and measures, or statistics and probability), in addition to overall mathematical achievement, is an important feature of the current study. Limited attention has been paid to variation in gendered achievement by mathematical content domain (i.e., in number work, algebra, geometry and measures, and statistics and probability) in the existing literature, which places considerable emphasis on overall mathematical achievement. This is an important area for investigation since mathematical performance has a tendency to be viewed as a monolith (Miner, 2019).

Northern Ireland provides an appropriate context for the current study since it is a jurisdiction that has attracted less attention in the global literature and also exhibits negligible gender differentials in mathematical achievement in large-scale international comparative studies such as PISA and TIMSS (Jerrim & Shure, 2016; Mullis et al., 2016; Wheater et al., 2013). Furthermore, Northern Ireland operates a system of academic selection, whereby children are selected for different types of post-primary schools, either grammar or non-grammar, on the basis of academic ability at 10 or 11 years of age. Grammar schools offer a more academically-oriented curriculum than non-grammar schools, which place a greater emphasis on the development of practical and vocational skills. However, it is noteworthy that students from lower socioeconomic backgrounds are less likely to attend grammar schools than their counterparts from more privileged backgrounds. For example, during the 2015–2016 school year, 13.9% of grammar school students were entitled to free school meals (which is considered to be a measure of socioeconomic advantage/disadvantage in Northern Ireland, with lower percentages indicating that students come from more affluent backgrounds), compared to 39.9% of non-grammar school students (Department of Education, 2016). Whilst grammar school students have been shown to significantly outperform their non-grammar school counterparts in international comparative studies of mathematical achievement (Jerrim & Shure, 2016), there is limited evidence in the literature pertaining to how gender differences in mathematical achievement vary by school type. The current study also addresses this gap in the current evidence base.

In the current research, multilevel models were used to analyse gender differences in overall mathematical achievement, and achievement in the various content domains of mathematics, based on results obtained in Northern Irish GCSE mathematics examinations. A multilevel regression analysis was also used to test for the effects of interactions of gender with type of post-primary school attended on overall mathematical achievement, and achievement in the different mathematical content domains. Given the dearth of studies that directly analyse high-stakes examination results in mathematics from a gender equity perspective using multilevel models, the current study makes a timely contribution to the literature pertaining to gender effects in mathematical achievement.

Historical Overview of Research on Gender and Mathematical Achievement

Traditionally, mathematics has been viewed as a discipline that is more accessible to men than women (Henrion, 1997). Early researchers in the field reported that there were no significant differences between boys' and girls' mathematical achievements in early primary school, but that differences in favour of boys began to emerge as children progressed in their education (e.g., Fennema, 1974). However, some inconsistencies were noted in the gender differentials, and the gender favoured, in different mathematical content domains and/or different countries (Hanna et al., 1990; Smith & Walker, 1988). Interestingly, the greatest differentials in favour of boys' mathematical achievements usually occurred amongst high-achieving students (Benbow & Stanley, 1980).

More recent research on gender effects in mathematical achievement, conducted during the early part of the twenty-first century, has broadly supported the findings reported by the pioneering researchers in the area. For example, in a study designed to investigate gender differences in the mathematical achievement of children aged nine, 12 or 15 years in Northern Taiwan, Chen and colleagues (2013) reported that gender had small to moderate effects on mathematical achievement, and that gender differences in favour of boys increased with age. These findings are corroborated by the work of Contini et al. (2017), who found that, after controlling for a variety of other background variables in an Italian context, boys perform better in mathematics than girls, and that the difference increases with age, with the greatest differentials for the highest-attaining students. These findings are consistent with those reported by Matteucci and Mignani (2011), who also investigated gender differences in mathematics for Italian students. Gender differentials in favour of boys' mathematical achievement have been highlighted by numerous researchers in relation to other national contexts (e.g., Dickerson et al., 2015; Leahey & Guo, 2001; Pargulski & Reynolds, 2017; Rodríguez-Planas & Nollenberger, 2018). However, it is important to observe that boys do not have a universal advantage in terms of mathematical achievement. This is exemplified by the fact Innabi and Dodeen (2018) concluded that 8th grade girls in Jordan outperformed boys in mathematics, but with the caveat that boys had a greater likelihood of correctly answering more demanding problems set in unfamiliar contexts, while girls were more likely to correctly answer routine problems. Gender differences in favour of boys have been reported in many studies involving PISA (Organisation for Economic Co-operation and Development (OECD, 2015, 2016), but the findings are more subtle in the case of TIMSS. This is illustrated by the fact no significant differences between boys' and girls' mathematics scores were recorded in 26 of the 39 countries participating in the TIMSS 2015 8th grade study (Mullis et al., 2016).

It is important to highlight that a very high proportion of previous research into gender differentials, including all studies cited thus far in the current section, was conducted using the results of low-stakes assessments of mathematical achievement, which had no direct consequences for study participants. There is less evidence pertaining to gender differences in relation to test results garnered from high-stakes assessments. Notable exceptions include Cox et al.'s (2004) study in an Australian context, and Zawistowska's (2017) study in a Polish context. Cox et al. (2004) reported that, on average, girls outperformed boys in almost all mathematical subjects for the majority of the years considered in the study, while Zawistowska (2017) concluded there were marginal gender differences in mathematical attainment, although boys demonstrated a pronounced advantage at the highest levels of performance.

Gender Differentials in International Large-Scale Assessments: Limitations of Low-Stakes Testing

The mathematical achievement of students from different countries is currently assessed in two major international comparative studies: the quadrennial TIMSS, which was inaugurated by the International Association for the Evaluation of Educational Achievement (IEA) in 1995, and the triennial PISA, which was introduced by the OECD in 2000. Both tests are deemed to yield valid international comparisons of students' mathematical achievements, and to have the potential to inform the decision-making of mathematics education policymakers in participating jurisdictions (Dossey & Wu, 2013), although the robustness of their underpinning measurement model has been challenged by some researchers (e.g., Cantley, 2015, 2017, 2019; Dohn, 2007; Goldstein, 2004). TIMSS test items are focused on assessing samples of 4th grade and 8th grade students' mastery of the content of the school mathematics curriculum, while PISA items attempt to assess the mathematical problem-solving skills of a sample of 15 year-olds in the context of real life scenarios. According to Dossey and Wu (2013), "although PISA's results provide a picture of students' capabilities, they provide less direct relationships to the schooling students have received" (p. 1013). However, they acknowledge that, despite the dubious link between schooling and PISA outcomes, the results "may provide a better picture of the future capabilities of nations' students to cope with everyday applications of mathematics" (p. 1013).

From a gender equity perspective, it is therefore concerning that the PISA 2012 study, which focused on mathematics, revealed that boys' mean mathematics score exceeded that of girls by 11 points. Boys outperformed girls in 38 of the 65 participating jurisdictions, whereas girls performed better than boys in just five countries (OECD, 2014, 2015). However, although the gender difference was significant at the 5% level, it is noteworthy that the male advantage was small (d = .109), and likely to be of limited practical significance. Furthermore, it is noteworthy that gender differences in PISA mathematical achievement vary considerably by country (Else-Quest et al., 2010). Indeed, in a number of jurisdictions, such as Northern Ireland, there were very small differences between boys' and girls' mathematical achievement levels in PISA 2012, and the associated effect size of d = .118 confirms that boys' scores only exceeded those for girls by a small margin (OECD, 2015; Wheater et al., 2013).

DeMars et al. (2013) noted that girls tend to expend greater effort than boys when confronted with low-stakes tests to the extent that "it is possible to observe sizeable gender differences in performance on low-stakes assessments partly or fully due to gender differences in test-taking motivation" (p. 79). Furthermore, Barry and colleagues (2010) posit that students are likely to invest greater effort in high-stakes testing situations where the test result has tangible personal consequences which, according to DeMars et al.'s (2013) argument, may influence the observed gender differentials in mathematical achievement for studies such as TIMSS and PISA. This stance is also advocated by Guez et al. (2020), who argue that gender differences in academic achievement are modulated by the stakes associated with the assessment used. The high-stakes test results that feature in the current research permitted an investigation into the import of these concerns.

Country-Specific Factors That may Potentially Influence Gender Differentials in Mathematical Achievement

In addition to differential performance by content domain, a number of other factors have been linked to gender differences in mathematical achievement. For example, gender social norms in a particular country have been shown to influence gender differentials in mathematical achievement (Rodríguez-Planas & Nollenberger, 2018). Ayalon and Livneh's (2013) research revealed that countries with greater degrees of standardisation in terms of their curriculum, and where the curriculum is governed via the use of national examinations, have lower gender differences in mathematical achievement, although no evidence was found of a relationship between such standardisation measures and a country's average mathematical achievement. However, Leder and Forgasz (2018) stressed that the design of assessment instruments, including the response format employed (e.g., free response or multiple choice), is an important contributory factor to gender effects in mathematical achievement.

The jurisdiction where the current research was conducted (Northern Ireland) operates a bipartite, selective education system, whereby some students complete their post-primary education in more academically-oriented grammar schools, with the remainder attending non-selective post-primary schools, which cater for students of all abilities. The use of academic selection is a highly contentious issue in Northern Ireland. Although students from lower socioeconomic backgrounds are less likely to gain admission to grammar schools (Gardner, 2016), these schools remain popular with many parents. This is particularly worrying since there are high levels of social disadvantage through poverty in many areas of Northern Ireland, thus meaning that academic selection perpetuates social class divisions (Gardner, 2016). Although the available data in the current study did not permit an analysis of the interaction of individual students' socioeconomic statuses and genders on mathematical achievement, it did facilitate an analysis of gender differentials in achievement by school type (grammar versus non-grammar), which addresses a gap in the research evidence in its own right.

Aims of Current Study and Research Questions

The aim of the current study was to examine gender differentials in both overall mathematical achievement, and achievement in various content domains (number, algebra, geometry and measures, and statistics and probability). This aim was addressed by examining a sample of Northern Irish students' total scores and domain-specific scores derived from item-level scores in GCSE mathematics. Furthermore, the data were analysed to test for the effects of interactions of gender with school type on both overall and domain-specific achievement.

The specific research questions addressed in the current study are as follows:

1. What is the magnitude of the gender difference for students' overall achievement in a high-stakes Northern Irish mathematics examination?
2. What are the magnitudes of the gender differences for students' achievement in the different content domains of a high-stakes Northern Irish mathematics examination?
3. How do the magnitudes of gender differences in mathematical achievement vary by school type for a high-stakes Northern Irish mathematics examination?

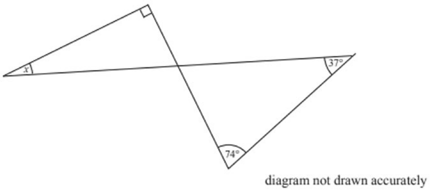
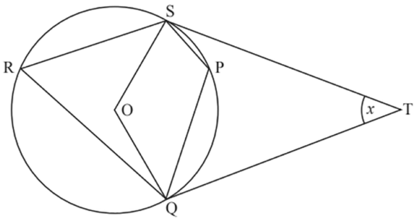
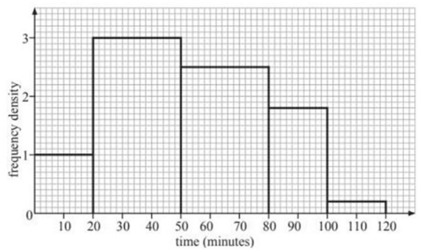
Methodology

The Current Study

The current study involved analysis of the results obtained by a sample of students in the summer 2016 Northern Ireland Council for the Curriculum, Examinations and Assessment (NICCEA) GCSE mathematics examinations. The NICCEA GCSE mathematics specification at that time was unitised and had two tiers of entry: foundation (usually entered by lower attaining students) and higher tier (usually entered by higher attaining students). A choice of units, with different levels of challenge, was available at each tier to cater for the wide range of mathematical abilities of candidates taking the tier. Scores from two units were used to determine the GCSE mathematics grade at certification stage. Each unit contained questions designed to assess students' knowledge, understanding, and problem-solving skills in relation to the four content domains: number, algebra, geometry and measures, statistics and probability. Unit assessments were marked in raw marks but, since candidates could take unit assessments in different examination series, raw marks were adjusted to *uniform marks* to compensate for any variation in difficulty of the assessments between series. This ensured that candidates who demonstrated the same level of achievement in a unit taken in different examination series finished up with the same uniform mark. Uniform marks from two assessment units were added to obtain a total uniform mark which determined the final GCSE mathematics grade.

NICCEA supplied the authors with anonymised item level raw marks, together with the total uniform mark, gender, and type of school attended (grammar or non-grammar) for each candidate who took a GCSE mathematics assessment unit in the summer 2016 examination series. Recall that grammar schools provide an academically-oriented curriculum, whereas non-grammar schools provide a practically/vocationally-oriented curriculum. Unfortunately, additional

Table 1 Sample Questions for each Content Domain at Foundation and Higher Tiers

Content domain	Foundation tier	Higher tier
Number	Last week a dentist noted that, of all her treatments, $\frac{1}{3}$ were fillings, $\frac{1}{4}$ were extractions, $\frac{1}{8}$ were denture treatment and the rest were cleaning. What fraction were cleaning?	A special offer shampoo bottle contains 20% extra. It contains 900 ml of shampoo. How much shampoo was in the original bottle?
Algebra	Expand $3(5 - y)$	Expand $(2x - 3)(3x + 4)$
Geometry and measures	Calculate the size of the angle marked x 	In the diagram, O is the centre of the circle P, Q, R and S are points on the circumference of the circle ST and QT are tangents to the circle Angle STQ = x Work out the size of angle SPQ in terms of x 
Statistics and probability	The lengths of twigs measured to the nearest tenth of a centimetre are given below. 4.3 4.7 2.9 1.0 5.8 4.2 3.6 1.9 2.7 3.0 2.6 3.7 4.3 2.7 2.8 Find the median of the lengths.	The histogram illustrates how much time drivers took on a particular journey.  Calculate an estimate for the mean time.

Northern Ireland Council for the Curriculum, Examinations and Assessment summer 2016 GCSE mathematics examination papers for units T2 (foundation tier) and T4 (higher tier) sociodemographic information pertaining to candidates was not available in the dataset supplied by NICCEA. The research entailed analysis of the results obtained by candidates taking the most popular combination of assessment units at foundation tier for whom certification of the qualification was also requested in summer 2016 (units T2 and T5), and likewise at higher tier (units T4 and T6). Further details pertaining to unit content are available from NICCEA (2017).

Initially, each question, at both foundation and higher tiers, was independently classified into one of the four content domains (number, algebra, geometry and measures, or statistics and probability) by both authors. Table 1 presents one sample question for each content domain at both foundation and higher tiers. Initial coding of questions resulted in 80% agreement between the authors at foundation tier, and 87.5% agreement at higher tier, regarding the distribution of the 200 raw marks available at each tier across the four content domains. Differences were resolved through discussion and agreement. Based on the supplied item level raw marks, the raw percentage scores obtained by candidates in each of the four content domains were then calculated.

The research was conducted in line with the research governance regulations of Queen's University Belfast, and the current study was approved by the research ethics committee of the university.

Participants and Sampling

The participants consisted of those NICCEA GCSE mathematics candidates for whom certification was requested in summer 2016, but who also took both assessment units that contributed to the certification (at foundation or higher tier, as appropriate) in the summer 2016 examination series. This was to allow for direct comparability of student achievement in each of the four mathematics content domains since the candidates concerned would have been tasked with answering exactly the same questions. Although certificates were issued for 16,351 candidates in summer 2016, the requirement for candidates to have taken both contributory assessment units in that series meant the sample consisted of 1,118 candidates from 113 schools at foundation tier (51.8% boys, 48.2% girls), and 2,762 candidates from 91 schools at higher tier (50.5% boys, 49.5% girls).

Variables

The variables included in the analysis are detailed below. The same variables featured in the analysis of both foundation and higher tier results.

Standardised total uniform score (z_{total}). The total uniform score used to determine GCSE grade at certification, standardised ($M = 0$, $SD = 1$) for the sampled participants at either foundation or higher tier, as appropriate.

Gender. Girls [Women] (0) or Boys [Men] (1).

School type (*SchType*). Non-grammar (0) or Grammar (1).

Standardised number score (z_{num}). Score in the number content domain, standardised ($M = 0$, $SD = 1$) for the sampled participants at either foundation or higher tier, as appropriate.

Standardised algebra score (z_{alg}). Score in the algebra content domain, standardised ($M = 0$, $SD = 1$) for the sampled participants at either foundation or higher tier, as appropriate.

Standardised geometry and measures score (z_{gm}). Score in the geometry and measures content domain, standardised ($M = 0$, $SD = 1$) for the sampled participants at either foundation or higher tier, as appropriate.

Standardised statistics and probability score (z_{sp}). Score in the statistics and probability content domain, standardised ($M = 0$, $SD = 1$) for the sampled participants at either foundation or higher tier, as appropriate.

Statistical Analysis

Descriptive statistics were used to provide an overview of the foundation and higher tier samples, including the means and standard deviations of boys' and girls' standardised total uniform scores, and standardised content domain-specific scores, for all school types, and also for grammar and non-grammar schools. Since the samples were clustered within schools, scores for students within the same school were likely to have more in common than scores of students from other schools (Cohen et al., 2011). Therefore, two-level multilevel regression models, with student at level one and school at level two, were used to address the three research questions. Multilevel modelling takes cognisance of the clustering of students within schools, and facilitates estimation of the variance in the dependent variable that is due to differences both within and between different schools (Field, 2013; Robson & Pevalin, 2016). Such an approach ensured that the standard errors of the regression coefficients were not underestimated (Woltman et al., 2012).

The first two research questions were addressed by estimating, for each tier (foundation and higher), the following two-level random intercept models:

$$z_{ij} = \beta_0 + u_{0j} + \varepsilon_{ij} \tag{1}$$

$$z_{ij} = \beta_0 + \beta_1 * Gender_{ij} + u_{0j} + \varepsilon_{ij} \tag{2}$$

where z_{ij} denotes the standardised score (z_{total} for research question one, and z_{num}, z_{alg}, z_{gm} or z_{sp} for research question two) for student i within school j; β_0 is the overall mean of the standardised score across all schools; u_{0j} is the random error at level two; ε_{ij} is the level one residual; and β_1 is the

Table 2 Mathematical Achievement by Gender and School Type at Foundation Tier

		All school types			Non-Grammar			Grammar		
Variable	Gender	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Overall mathematical achievement	Boys	579	.063	1.033	451	-.110	1.049	128	.671	.694
	Girls	539	-.067	.959	479	.172	.950	60	.767	.536
Number	Boys	579	.103	1.047	451	-.067	1.055	128	.704	.758
	Girls	539	-.111	.935	479	-.214	.914	60	.714	.659
Algebra	Boys	579	-.045	1.012	451	-.201	1.028	128	.507	.727
	Girls	539	.048	.985	479	-.044	.988	60	.785	.574
Geometry and measures	Boys	579	.131	1.044	451	.003	1.041	128	.585	.923
	Girls	539	-.141	.931	479	-.231	.913	60	.575	.755
Statistics and probability	Boys	579	-.005	1.038	451	-.188	1.036	128	.641	.753
	Girls	539	.005	.958	479	-.069	.951	60	.602	.799

N=1,118. Variables have been standardised for the sample at foundation tier. Data derived from information supplied by Northern Ireland Council for the Curriculum, Examinations and Assessment

$Gender_{ij}$ slope coefficient. A significant gender differential in mathematical achievement is indicated by a non-zero (significant) β_1 coefficient. Since the outcome scores are standardised, Lorah (2018) suggests that the β_1 coefficients provide an appropriate measure of the effect size after controlling for nesting within schools. It has been proposed by some researchers that Cohen's *d*, which measures the standardised mean difference between the two groups defined by a dichotomous variable (such as *Gender*) is an appropriate measure of effect size for a dichotomous covariate in a multilevel regression model (Snijders & Bosker, 2012). However, Cohen's *d* simply quantifies the relationship between the dependent variable and the dichotomous covariate without controlling for clustering effects at level two of the model. Cohen's *d* is thus an inappropriate measure of effect size for data with a hierarchical structure, such as those in the current study. Instead, the β_1 coefficient in the model indicates the expected number of standard deviation increases in mathematical achievement associated with being male, but controlling for school effects, and therefore represents a more robust measure of effect size.

The third research question was addressed by estimating, for each tier (foundation and higher), the following two-level random intercept models:

$$z_{ij} = \beta_0 + \beta_1 * Gender_{ij} + \beta_2 * SchType_j + u_{0j} + \varepsilon_{ij} \tag{3}$$

$$\begin{aligned} z_{ij} =& \beta_0 + \beta_1 * Gender_{ij} + \beta_2 * SchType_j + \beta_3 \\ & * SchType_j * Gender_{ij} + u_{0j} + \varepsilon_{ij} \end{aligned} \tag{4}$$

where z_{ij} denotes the standardised score (z_{total}, z_{num}, z_{alg}, z_{gm} or z_{sp} as appropriate) for student i within school j; β_0 is the overall mean of the standardised score across all schools; u_{0j} is the random error at level two; ε_{ij} is the level one residual; β_1 is the $Gender_{ij}$ slope coefficient; β_2 is the $SchType_j$ slope

Table 3 Mathematical Achievement by Gender and School Type at Higher Tier

		All school types			Non-Grammar			Grammar		
Variable	Gender	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Overall mathematical achievement	Boys	1,395	.054	.974	91	-.555	1.233	1,304	.096	.940
	Girls	1,367	-.055	1.023	91	-.727	1.276	1,276	-.007	.986
Number	Boys	1,395	.077	.956	91	-.406	1.247	1,304	.111	.923
	Girls	1,367	.079	1.038	91	.662	1.331	1,276	-.037	1.001
Algebra	Boys	1,395	.018	1.010	91	-.557	1.177	1,304	.059	.985
	Girls	1,367	-.019	.990	91	-.479	.944	1,276	.014	.985
Geometry and measures	Boys	1,395	.079	.964	91	-.432	1.116	1,304	.114	.943
	Girls	1,367	-.080	1.030	91	-.674	1.020	1,276	-.038	1.018
Statistics and probability	Boys	1,395	.037	.979	91	-.415	1.181	1,304	.069	.956
	Girls	1,367	-.038	1.020	91	-.642	1.129	1,276	.005	.998

N=2,762. Variables have been standardised for the sample at higher tier. Data derived from information supplied by Northern Ireland Council for the Curriculum, Examinations and Assessment coefficient; and β_3 is the slope coefficient for the interaction of *Gender* and *SchType*.

All statistical calculations were performed using Stata version 14.

Results

Descriptive Statistics

Tables 2 and 3 summarise, for foundation and higher tiers respectively, means and standard deviations for girls' and boys' overall mathematical achievement and achievement in each of the four content domains (number, algebra, geometry and measures, statistics and probability). However, because of the hierarchical structure of the dataset, the magnitudes of the gender differentials in achievement are more accurately assessed in the subsequent sections, which present the results of multilevel regression analyses that control for school effects.

Multilevel Assessment of Gender Effects in Overall and Domain-Specific Mathematical Achievement

A two-level multilevel analysis (with student at level one and school at level two) was performed to assess gender differentials in overall and content domain-specific achievement at both foundation and higher tiers, while controlling for clustering effects within schools. Initially, for each tier, a null model with no covariates, represented by Eq. (1), was estimated for each of the five achievement-related variables. The null model permitted investigation of whether there was evidence to justify the existence of random intercepts for the grouping variable at level two (school), and therefore whether multilevel modelling was necessary. Gender was then introduced as a level one explanatory variable to estimate the multilevel model represented by Eq. (2) for each of the five achievement variables. The estimated regression coefficients (with standard errors in parentheses) are presented in Table 4 for foundation tier and Table 5 for higher tier. Level one and level two variances, together with the intraclass correlation coefficient (ICC) and log likelihood, are also included for each model in Tables 4 and 5.

The null models for overall mathematical achievement, at both foundation and higher tiers, indicated that substantial proportions of the variation in achievement were explained by differences between schools, 25.1% at foundation tier and 28.0% at higher tier, as revealed by the ICCs of .251 and .280 respectively. This confirmed that multilevel modelling was appropriate when assessing the

Table 4 Multilevel Models to Assess Gender Effects in Mathematical Achievement at Foundation Tier

	Overall achievement		Number		Algebra		Geometry and measures		Statistics and probability	
	Null	Null + Gender	Null	Null + Gender	Null	Null + Gender	Null	Null + Gender	Null	Null + Gender
Intercept	.094 (.062)	.080 (.070)	.109 (.062)	.051 (.069)	.079 (.058)	.169* (.067)	.055 (.056)	-.034 (.064)	.046 (.059)	.097 (.068)
Gender [reference: girls]		.027 (.062)		.111 (.062)		-.173** (.063)		.170** (.063)		-.098 (.063)
Level 1 variance	.791	.791	.801	.800	.827	.820	.829	.826	.827	.823
Level 2 variance	.265	.264	.258	.251	.214	.219	.192	.182	.218	.225
ICC	.251	.250	.244	.239	.206	.211	.188	.181	.209	.214
Log likelihood	-1521.078	-1520.980	-1526.919	-1525.347	-1537.578	-1533.815	-1535.606	-1531.934	-1538.078	-1536.884

N = 1,118. Data derived from information supplied by Northern Ireland Council for the Curriculum, Examinations and Assessment

ICC = intraclass correlation coefficient

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table 5 Multilevel Models to Assess Gender Effects in Mathematical Achievement at Higher Tier

	Overall achievement		Number		Algebra		Geometry and measures		Statistics and probability	
	Null	Null + Gender	Null	Null + Gender	Null	Null + Gender	Null	Null + Gender	Null	Null + Gender
Intercept	-.221** (.072)	-.249** (.074)	-.141* (.058)	-.190** (.061)	-.184** (.060)	-.167** (.064)	-.146* (.060)	-.200** (.063)	-.154** (.058)	-.188** (.062)
Gender [reference: girls]		.057 (.041)		.098* (.042)		-.035 (.041)		.108** (.041)		.070 (.042)
Level 1 variance	.840	.839	.887	.885	.870	.870	.873	.871	.894	.893
Level 2 variance	.327	.327	.188	.184	.211	.211	.208	.206	.188	.188
ICC	.280	.280	.175	.172	.195	.196	.193	.191	.174	.174
Log likelihood	-3761.233	-3760.264	-3818.240	-3815.469	-3795.748	-3795.389	-3799.895	-3796.519	-3828.697	-3827.317

N=2,762. Data derived from information supplied by Northern Ireland Council for the Curriculum, Examinations and Assessment

ICC intraclass correlation coefficient

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

influence of gender on overall achievement at both tiers. However, when gender was added as a level one explanatory variable to both the foundation and higher tier models, there was no appreciable change in either the ICCs or the log likelihood functions, which suggested that gender had minimal influence in explaining differences in overall achievement. This is further reinforced by the fact the gender coefficients of .027 and .057 for foundation and higher tiers respectively were not significant, even at the 5% level. The small, positive magnitudes of these effect sizes indicate that, for both tiers, boys scored very marginally higher than girls on average, but the differentials are negligible for practical purposes. On average, boys scored .027 of a standard deviation higher than girls at foundation tier, and .057 of a standard deviation at higher tier.

At foundation tier, only algebra and geometry and measures, had significant gender coefficients (at the 1% level) of -.173 and .170 respectively. These values showed that, at foundation tier, boys on average scored .173 of a standard deviation lower than girls in algebra, but .170 of a standard deviation higher in geometry and measures. Although these differentials are greater in magnitude than those for overall achievement, both are relatively small. At higher tier, only number and geometry and measures revealed significant gender differentials, with small effect sizes of .098 and .108 respectively. These values indicate that, at higher tier, boys on average scored .098 of a standard deviation higher than girls in the number domain and .108 of a standard deviation higher in geometry and measures.

Effects of School Type and Gender on Overall and Domain-Specific Mathematical Achievement

To investigate the effects of school type and gender on overall and domain-specific mathematical achievement at each tier (foundation and higher), two-level random intercept models (again with student at level one and school at level two) were fitted with the standardised score for mathematical achievement as the dependent variable. Initially, gender and school type were used as covariates, but the interaction term between gender and school type was then introduced. The results of these multilevel analyses are presented in Tables 6 and 7.

At both foundation and higher tiers, school type is a highly significant predictor (at the .1% level) of overall and domain-specific mathematical achievement, with grammar school students performing significantly better than their non-grammar school counterparts. This corroborates Jerrim and Shure's (2016) evidence of achievement differentials between grammar and non-grammar school students in Northern Ireland. It is also noteworthy that the majority of

Table 6 Multilevel Models to Assess Gender and School Type Effects in Mathematical Achievement at Foundation Tier

	Overall achievement		Number		Algebra		Geometry and measures		Statistics and probability	
	Gender + School type	Gender + School type + Interaction term	Gender + School type	Gender + School type + Interaction term	Gender + School type	Gender + School type + Interaction term	Gender + School type	Gender + School type + Interaction term	Gender + School type	Gender + School type + Interaction term
Intercept	-.125* (.063)	-.136* (.064)	-.158* (.062)	-.169** (.063)	-.014 (.061)	-.023 (.062)	-.185** (.060)	-.200** (.061)	-.074 (.061)	-.063 (.062)
Gender [reference: girls]	.018 (.061)	.039 (.065)	.103 (.061)	.126 (.065)	-.187** (.061)	-.168** (.066)	.160** (.061)	.192** (.066)	-.103 (.062)	-.124 (.066)
School type [reference: non-grammar]	.836*** (.118)	.933*** (.159)	.831*** (.115)	.935*** (.157)	.774*** (.115)	.858*** (.157)	.653*** (.112)	.794*** (.156)	.736*** (.114)	.641*** (.157)
Gender * School Type		-.164 (.181)		-.174 (.181)		-.141 (.182)		-.237 (.181)		.160 (.182)
Level 1 variance	.795	.795	.805	.804	.822	.821	.829	.828	.832	.831
Level 2 variance	.124	.124	.113	.113	.109	.108	.100	.099	.104	.104
ICC	.135	.135	.123	.123	.117	.117	.107	.107	.112	.111
Log likelihood	-1501.278	-1500.867	-1504.909	-1504.443	-1515.280	-1514.979	-1517.944	-1517.089	-1520.771	-1520.386

N = 1,118. Data derived from information supplied by Northern Ireland Council for the Curriculum, Examinations and Assessment

ICC intraclass correlation coefficient

**p* < .05; ***p* < .01; ****p* < .001

Table 7 Multilevel Models to Assess Gender and School Type Effects in Mathematical Achievement at Higher Tier

	Overall achievement		Number		Algebra		Geometry and measures		Statistics and probability	
	Gender + School type	Gender + School type + Interaction term	Gender + School type	Gender + School type + Interaction term	Gender + School type	Gender + School type + Interaction term	Gender + School type	Gender + School type + Interaction term	Gender + School type	Gender + School type + Interaction term
Intercept	-.717*** (.127)	-.806*** (.144)	-.592*** (.111)	-.685*** (.132)	-.577*** (.117)	-.582*** (.135)	-.593*** (.114)	-.681*** (.134)	-.605*** (.112)	-.691*** (.132)
Gender [reference: girls]	.057 (.041)	.238 (.142)	.100* (.041)	.287* (.144)	-.034 (.041)	-.023 (.144)	.109** (.041)	.287* (.144)	.069 (.042)	.242 (.145)
School type [reference: non-grammar]	.650*** (.144)	.746*** (.162)	.527*** (.123)	.628*** (.145)	.549*** (.130)	.555*** (.150)	.523*** (.127)	.618*** (.148)	.553*** (.123)	.647*** (.145)
Gender * School Type		-.197 (.149)		-.203 (.151)		-.012 (.150)		-.194 (.150)		-.188 (.151)
Level 1 variance	.841	.840	.886	.885	.870	.870	.872	.872	.894	.893
Level 2 variance	.224	.229	.128	.131	.159	.159	.145	.148	.127	.129
ICC	.210	.214	.126	.128	.154	.154	.142	.146	.124	.126
Log likelihood	-3751.397	-3750.528	-3807.429	-3806.522	-3787.291	-3787.288	-3789.185	-3788.352	-3818.431	-3817.657

N = 2,762. Data derived from information supplied by Northern Ireland Council for the Curriculum, Examinations and Assessment

ICC intraclass correlation coefficient

$*p < .05$; $**p < .01$; $***p < .001$

students at foundation tier attended non-grammar schools, while the majority at higher tier attended grammar schools.

At foundation tier, the ICCs and log likelihoods in Table 6 indicate that the addition of school type as a level two explanatory variable in the model for overall mathematical achievement further reduced the variance between schools and also improved upon the fit of the model, relative to the null plus gender model presented in Table 4. However, the gender coefficient was still not significant, nor was the coefficient of the gender x school type interaction term, thus implying that school type did not have a bearing on the gender differential in overall achievement at foundation tier. Gender differentials for the four content domains at foundation tier were also in line with those presented in Table 4. It is notable that the coefficients of the gender x school type interaction terms were not significant, thus confirming that school type did not significantly influence gender differentials within the four content domains. At higher tier, the addition of school type as a level two explanatory variable in the models yielded similar conclusions about the statistical significance of the gender coefficients to those obtained for the null plus gender models presented in Table 5. As for foundation tier, none of the coefficients for the gender x school type interaction terms were significant, thus suggesting that school type did not have an impact on gender differentials at higher tier.

Discussion and Conclusions

Summary of Findings

What is the magnitude of the gender difference for students' overall achievement in a high-stakes Northern Irish mathematics examination? Findings from the current study provide evidence in support of the gender similarities hypothesis (Hyde, 2005, 2014, 2016). Specifically, the results of the multilevel analysis confirmed that the differences between boys' and girls' overall mathematical achievement in the high-stakes examination were close-to-zero for both foundation and higher tiers, with effect sizes of .027 and .057 respectively. This is consistent with other research focusing on gender differentials in overall mathematical achievement within a Northern Irish context, albeit research that was not undertaken using the results obtained in high-stakes examinations (Jerrim & Shure, 2016; Mullis et al., 2016; Wheater et al., 2013).

What are the magnitudes of the gender differences for students' achievement in the different content domains of a high-stakes Northern Irish mathematics examination? The multilevel analysis indicated that gender differentials in the foundation tier examination were significant at the 1% level for both algebra and geometry and measures, with relatively small effect sizes of -.173 and .170 respectively. These demonstrated a small female advantage in algebra and a similar small male advantage in relation to geometry and measures. Interestingly, the multilevel analysis also highlighted a very small differential in favour of girls for algebra at higher tier, although it did not attain statistical significance, even at the 5% level. At higher tier, only the number and geometry and measures content domains exhibited significant gender differentials in favour of boys, with small effect sizes of .098 and .108 respectively. However, it is important to note that all of the effect sizes are small and offer evidence to support the gender similarities hypothesis in respect of domain-specific mathematical achievement (Hyde, 2005, 2014, 2016).

The existence of significant, but small, gender effects in relation to content domain-specific achievement contradicts some existing research that was conducted using international large-scale assessments of mathematical achievement in Northern Ireland (e.g., Wheater et al., 2013). However, the findings resonate with the conclusions reported by some other scholars internationally. For example, Leahey and Guo (2001) and Liu and Wilson (2009) both concluded that boys demonstrated significantly higher achievement than girls in geometry. Furthermore, the female advantage in algebraic achievement identified in the current study aligns with the findings of Louis and Mistele (2012), who reported that, in their US-focused research, algebra was the only mathematical content domain where girls outperformed boys. Therefore, there are slight disparities between the conclusions drawn from recent research into Northern Irish gender differentials in content domain-specific mathematical achievement that was conducted using comparative international assessment data (e.g., from PISA) and the findings of the current study. The disparity is particularly apparent in relation to the gender differential in favour of males for achievement in geometry and measures that featured at both tiers in the current study, but it is noteworthy that all of the effect sizes are small.

How do the magnitudes of gender differences in mathematical achievement vary by school type for a high-stakes Northern Irish mathematics examination? Results of the multilevel analysis highlight that, while grammar school students consistently outperformed their non-grammar school counterparts at both tiers of entry (with large effect sizes), school type did not have a bearing on gender differentials for either overall or domain-specific mathematical achievement. However, it is important to note that the majority of students at foundation tier attended non-grammar schools, while the majority at higher tier attended grammar schools, thus suggesting a potential equity issue pertaining to tier of entry.

Limitations and Future Research Directions

Although the current study involved multilevel analysis of high-stakes mathematics examination results of 3,880 students (1,118 of whom took foundation tier, and 2,762 of whom took higher tier examinations), this represented just 23.7% of the 16,351 candidates who received a NICCEA GCSE mathematics certificate in summer 2016. The sample was chosen to include only those candidates who took the most popular combination of assessment units, at both foundation and higher tiers, for whom certification of the qualification was requested in summer 2016, and where both contributory assessment units were also taken during that examination series. Although this ensured that identical questions were attempted by all candidates within a particular tier, and thus permitted domain-specific achievement to be assessed more reliably, it does restrict the generalisability of the findings.

The current study also attempted to offer some insights into the similarities and differences in the conclusions that can be drawn from investigations into gender differentials in achievement using different data sources, namely the results of high-stakes examinations versus performance in low-stakes assessments. The credence that can be accorded to these insights is restricted by the use of different samples of students, although some reassurance can be taken from the relatively large sample used in the current study. Clearly, it would be preferable to undertake a study involving the analysis of high-stakes examination results for a sample of students who also participated in low stakes assessments such as PISA or TIMSS, so that more robust comparisons between the different approaches to assessing gender differentials could be established.

The current study focused on overall and content domain-specific achievement. However, there is scope to classify the various items on the assessment units into cognitive domains (e.g., knowledge, application, reasoning, problem-solving) rather than content domains, to facilitate an analysis of gender differentials in cognitive domain-specific achievement. In addition, a more fine-grained analysis of domain-specific achievement, coupled with scrutiny of the actual assessment items, may offer useful insights into assessment practices that inadvertently lead to gender differentials in achievement. For example, it is conceivable that some assessment practices could be experienced differently by boys [men] and girls [women], and that these different experiences may actually precipitate gender differentials in achievement.

Finally, the lack of student sociodemographic information in the dataset supplied by NICCEA significantly limited the analysis that could be performed. Nevertheless, the analysis that was possible has contributed new knowledge to the field of gender differences in mathematical achievement.

Practice Implications

On the basis of the findings of the current research, coupled with the overview of previous studies outlined in the literature review, some recommendations can be proffered for mathematics education policymakers in Northern Ireland, which are also likely to have considerable import for policymakers in other international contexts. The results of the multilevel analysis of high-stakes mathematics examination scores suggest that there is gender equity in overall mathematical achievement in Northern Ireland. Given that this resonates with the findings of other recent studies, it is conceivable that the Northern Ireland education system, together with its mathematics curriculum and assessment arrangements are conducive to promoting gender equity in overall mathematical achievement. Therefore, consideration of Northern Ireland as a case study may serve as a useful starting point for other countries with an interest in promoting more gender-equitable outcomes in mathematical achievement.

Whilst there is gender equity in overall achievement, small but significant differences are apparent for achievement in some of the content domains. In particular, the gender equity in overall achievement masks a consistent, small male advantage in geometry and measures. This contradicts other recent findings on gender differentials for content domain-specific achievement in Northern Ireland (e.g., Wheater et al., 2013), but corroborates the conclusions reported by scholars working in other international contexts (e.g., Leahey & Guo, 2001; Liu & Wilson, 2009). This suggests that it is appropriate, both in Northern Ireland, and in other countries, to investigate interventions aimed at improving girls' achievement in geometry. For example, Meggiolaro (2018) demonstrated that the use of some ICT applications can lead to achievement gains in geometry for girls, and it is suggested that the potential of such applications should be explored as a vehicle for promoting higher levels of gender equity in mathematical learning outcomes, both in Northern Ireland and elsewhere. To this end, the dynamic geometry capabilities of software such as GeoGebra are likely to have beneficial roles to play in enhancing pedagogy.

The current study offers some useful insights into the appropriateness of using data from low-stakes assessments, such as PISA and TIMSS, rather than high-stakes assessments (with real consequences for students), to assess gender differentials in mathematical achievement. Both approaches led to similar conclusions about gender differentials in overall mathematical achievement in Northern Ireland, but a more subtle picture emerged in relation to content domain-specific achievement. Small, but significant, gender differentials in content domain achievement were apparent in the research conducted using high-stakes examination results, but it is important to note that the observed effect sizes are small. This indicates that studies using PISA data do offer useful insights into gender equity issues in mathematical achievement, despite concerns some researchers have raised about the utility of low-stakes tests for accurately assessing gender differentials in achievement. It has been suggested that the low-stakes nature of tests such as those used in PISA and TIMSS may potentially lead to differential gender effects in test-taking motivation, which could in turn precipitate gender differentials in achievement (Barry et al., 2010; DeMars et al., 2013; Eklöf, 2010; Guez et al., 2020). However, evidence from the current study, which entailed a multilevel analysis of high-stakes mathematics examination results from a jurisdiction where no significant gender differentials in mathematical achievement were apparent in PISA, suggests that these concerns are unfounded.

Conclusions

Gender stereotyping is prevalent in mathematics, which is often viewed as an academic discipline that is more appropriate for men than women due to inherent differences between the genders (Franceschini et al., 2014). The findings of the current research, using results from high-stakes summative assessments in mathematics, have demonstrated that there are negligible or small gender differentials in mathematical achievement. These findings support the gender similarities hypothesis, which asserts that males and females are similar on most, but not all, psychological variables (Hyde, 2005, 2014, 2016). No significant differences between boys and girls were found in relation to overall mathematical achievement, but small differences in favour of males were apparent in relation to achievement in the geometry and measures content domain. However, the magnitudes of the associated effect sizes mean that the differences are likely to be of limited practical significance. Therefore, it is imperative that damaging gender stereotyping in mathematics is dispensed with as a matter of urgency, and that improvements are made to mathematics pedagogy to allow girls to realise their mathematical potential. To this end, the current article has offered some suggestions regarding potential improvements to girls' learning experiences in geometry.

The current study also provided some insights into the viability of using data from low-stakes assessments such as PISA or TIMSS to accurately investigate gender differentials in achievement. Similar conclusions were drawn from the analysis of both these low-stakes assessments and high-stakes public examination results, thus indicating that research based on low-stakes assessments (such as PISA or TIMSS) does offer useful insights into gender differentials in achievement, thus contradicting the findings of Guez et al. (2020). Finally, findings from this study augment the evidence base pertaining to gender effects in mathematical achievement since they confirmed that school type, selective versus non-selective, did not have an effect on gender differences, an area where there was scant evidence in the existing literature.

The current study therefore offers useful insights for researchers and policymakers with an interest in enhancing gender equity in mathematical achievement, which is a highly desirable characteristic of any high-performing education system.

Funding No external funding was received for this project.

Data Availability Data is available from the first author upon request.

Declarations

Ethical Approval The study was approved by the research ethics committee of the university.

Research Involving Human and Animal Participants The research involved the analysis of anonymised data and was conducted in line with the research governance regulations of Queen's University Belfast.

Conflict of Interest There are no conflicts of interest.

References

Ayalon, H., & Livneh, I. (2013). Educational standardization and gender differences in mathematics achievement: A comparative study. *Social Science Research, 42*(2), 432–445. https://doi.org/10.1016/j.ssresearch.2012.10.001

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10*(4), 342–363. https://doi.org/10.1080/15305058.2010.508569

Benbow, C. P., & Stanley, J. (1980). Sex differences in mathematical ability: Facto or artefact? *Science, 210*(4475), 1262–1264. https://doi.org/10.1126/science.7434028

Brunner, M., Gogol, K. M., Sonnleitner, P., Keller, U., Krauss, S., & Preckel, F. (2013). Gender differences in the mean level, variability, and profile shape of student achievement: Results from 41 countries. *Intelligence, 41*(5), 378–395. https://doi.org/10.1016/j.intell.2013.05.009

Cantley, I. (2015). How secure is a Newtonian paradigm for psychological and educational measurement? *Theory & Psychology, 25*(1), 117–138. https://doi.org/10.1177/0959354314561141

Cantley, I. (2017). A quantum measurement paradigm for educational predicates: Implications for validity in educational measurement. *Educational Philosophy and Theory, 49*(4), 405–421. https://doi.org/10.1080/00131857.2015.1048668

Cantley, I. (2019). PISA and policy-borrowing: A philosophical perspective on their interplay in mathematics education. *Educational Philosophy and Theory, 51*(12), 1200–1215. https://doi.org/10.1080/00131857.2018.1523005

Chen, H.-Y., Chen, M.-F., Lee, Y.-S., Chen, H.-P., & Keith, T. Z. (2013). Gender reality regarding mathematic outcomes of students aged 9 to 15 years in Taiwan. *Learning and Individual Differences, 26*, 55–63. https://doi.org/10.1016/j.lindif.2013.04.009

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th ed.). Routledge.

Contini, D., Di Tommaso, M. L., & Mendolia, S. (2017). The gender gap in mathematics achievement: Evidence from Italian data. *Economics of Education Review, 58*, 32–42. https://doi.org/10.1016/j.econedurev.2017.03.001

Cox, P. J., Leder, G. C., & Forgasz, H. J. (2004). Victorian Certificate of Education: Mathematics, science and gender. *Australian Journal of Education, 48*(1), 27–46. https://doi.org/10.1177/000494410404800103

DeMars, C. E., Bashko, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, *8*(1), 69–82. https://files.eric.ed.gov/fulltext/EJ1062839.pdf.

Demir, I., Kiliç, S., & Ünal, H. (2010). Effects of students' and schools' characteristics on mathematics achievement: Findings from PISA 2006. *Procedia Social and Behavioral Sciences, 2*(2), 3099–3103. https://doi.org/10.1016/j.sbspro.2010.03.472

Department of Education. (2016). *Statistical bulletin 3/2016: Annual enrolments at schools and in funded pre-school education in Northern Ireland, 2015/16*. Retrieved from https://www.education-ni.gov.uk/sites/default/files/publications/education/Statistical%20Bulletin%201516%20-%20March%20%2820.07.16%20update%29.PDF.

Dickerson, A., McIntosh, S., & Valente, C. (2015). Do the maths: An analysis of the gender gap in mathematics in Africa. *Economics of Education Review, 46*, 1–22. https://doi.org/10.1016/j.econedurev.2015.02.005

Dohn, N. B. (2007). Knowledge and skills for PISA – assessing the assessment. *Journal of Philosophy of Education, 41*(1), 1–16. https://doi.org/10.1111/j.1467-9752.2007.00542.x

Dossey, J. A., & Wu, M. L. (2013). Implications of international studies for national and local policy in mathematics education. In M. A. (Ken) Clements, A. Bishop, C. Keitel-Kreidt, J. Kilpatrick, & F. K. S. Leung (Eds.), *Third international handbook of mathematics education* (pp. 1009–1042). Springer.

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice, 17*(4), 345–356. https://doi.org/10.1080/0969594x.2010.516569

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*(1), 103–127. https://doi.org/10.1037/a0018053

Fennema, E. (1974). Mathematics learning and the sexes: A review. *Journal for Research in Mathematics Education, 5*(3), 126–139. https://doi.org/10.2307/748949

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). SAGE Publishing.

Franceschini, G., Galli, S., Chiesi, F., & Primi, C. (2014). Implicit gender-math stereotype and women's susceptibility to stereotype threat and stereotype lift. *Learning and Individual Differences, 32*, 273–277. https://doi.org/10.1016/j.lindif.2014.03.020

Gardner, J. (2016). Education in Northern Ireland since the Good Friday Agreement: Kabuki theatre meets danse macabre. *Oxford Review of Education, 42*(3), 346–361. https://doi.org/10.1080/03054985.2016.1184869

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education: Principles, Policy & Practice, 11*(3), 319–330. https://doi.org/10.1080/0969594042000304618

Guez, A., Peyre, H., & Ramus, F. (2020). Sex differences in academic achievement are modulated by evaluation type. *Learning and Individual Differences, 83–84*, 101935. https://doi.org/10.1016/j.lindif.2020.101935

Hanna, G., Kündiger, E., & Larouche, C. (1990). Mathematical achievement of grade 12 girls in fifteen countries. In L. Burton (Ed.), *Gender and mathematics: An international perspective* (pp. 87–97). Cassell.

Henrion, C. (1997). *Women in mathematics: The addition of difference*. Indiana University Press.

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

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology, 65*(1), 373–398. https://doi.org/10.1146/annurev-psych-010213-115057

Hyde, J. S. (2016). Sex and cognition: Gender and cognitive functions. *Current Opinion in Neurobiology, 38*, 53–56. https://doi.org/10.1016/j.conb.2016.02.007

Innabi, H., & Dodeen, H. (2018). Gender differences in mathematics achievement in Jordan: A differential item functioning analysis of the 2015 TIMSS. *School Science and Mathematics, 118*(3–4), 127–137. https://doi.org/10.1111/ssm.12269

Jerrim, J., & Shure, N. (2016). *Achievement of 15-year-olds in Northern Ireland: PISA 2015 national report*. UCL Institute of Education. https://www.education-ni.gov.uk/sites/default/files/publications/education/ACHIEVEMENT%20OF%2015%20YEAR-OLDS%20IN%20NORTHERN%20IRELAND%20PISA%202015%20NATIONAL%20REPORT..PDF.

Leahey, E., & Guo, G. (2001). Gender differences in mathematical trajectories. *Social Forces, 80*(2), 713–732. https://doi.org/10.1353/sof.2001.0102

Leder, G. C., & Forgasz, H. J. (2018). Measuring who counts: Gender and mathematics assessment. *ZDM, 50*(4), 687–697. https://doi.org/10.1007/s11858-018-0939-z

Li, M., Zhang, Y., Liu, H., & Hao, Y. (2018). Gender differences in mathematics achievement in Beijing: A meta-analysis. *British Journal of Educational Psychology, 88*(4), 566–583. https://doi.org/10.1111/bjep.12203

Lindberg, S. M., Hyde, J. S., Peterson, 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

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education, 22*(2), 164–184. https://doi.org/10.1080/08957340902754635

Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education, 6*(1), 1–11. https://doi.org/10.1186/s40536-018-0061-2

Louis, R. A., & Mistele, J. M. (2012). The differences in scores and self-efficacy by student gender in mathematics and science. *International Journal of Science and Mathematics Education, 10*(5), 1163–1190. https://doi.org/10.1007/s10763-011-9325-9

Matteucci, M., & Mignani, S. (2011). Gender differences in performance in mathematics at the end of lower secondary school in Italy. *Learning and Individual Differences, 21*(5), 543–548. https://doi.org/10.1016/j.lindif.2011.03.001

McGraw, R., Lubienski, S. T., & Strutchens, M. E. (2006). A closer look at gender in NAEP mathematics achievement and affect data: Intersections with achievement, race/ethnicity, and socioeconomic status. *Journal for Research in Mathematics Education*, *37*(2), 129–150. https://www.jstor.org/stable/30034845.

Meggiolaro, S. (2018). Information and communication technologies use, gender and mathematics achievement: Evidence from Italy. *Social Psychology of Education: An International Journal, 21*(2), 497–516. https://doi.org/10.1007/s11218-017-9425-7

Miner, M. A. (2019). Unpacking the monolith: Intersecting gender and citizenship status in STEM graduate education. *International Journal of Sociology and Social Policy, 39*(9–10), 661–679. https://doi.org/10.1108/ijssp-05-2019-0101

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Boston College, Trends in International Mathematics and Science Study (TIMSS) & Progress in International Reading Literacy Study (PIRLS) International Study Center. http://timssandpirls.bc.edu/timss2015/international-results/.

Nix, S., & Perez-Felkner, L. (2019). Difficulty orientations, gender, and race/ethnicity: An intersectional analysis of pathways to STEM degrees. *Social Sciences, 8*(2), 43. https://doi.org/10.3390/socsci8020043

Nollenberger, N., Rodríguez-Planas, N., & Sevilla, A. (2016). The math gender gap: The role of culture. *American Economic Review, 106*(5), 257–261. https://doi.org/10.1257/aer.p20161121

Northern Ireland Council for the Curriculum, Examinations and Assessment (2017). *GCSE Mathematics (2017)*. https://ccea.org.uk/key-stage-4/gcse/subjects/gcse-mathematics-2017

Organisation for Economic Co-operation and Development (2014). *PISA 2012 results: What students know and can do – Student performance in mathematics, reading and science (volume I, revised edition, February 2014)*. OECD. https://www.oecd.org/pisa/keyfindings/pisa-2012-results-volume-I.pdf.

Organisation for Economic Co-operation and Development (2015). *The ABC of gender equality in education: Aptitude, behaviour, confidence*. OECD. https://www.oecd.org/pisa/keyfindings/pisa-2012-results-gender-eng.pdf.

Organisation for Economic Co-operation and Development (2016). *PISA 2015 results (volume I): Excellence and equity in education*. OECD. https://www.oecd-ilibrary.org/docserver/9789264266490-en.pdf?expires=1625054287&id=id&accname=guest&checksum=90800465380342A8017AA703D2C4803F.

Pargulski, J. R., & Reynolds, M. R. (2017). Sex differences in achievement: Distributions matter. *Personality and Individual Differences, 104*, 272–278. https://doi.org/10.1016/j.paid.2016.08.016

Perez-Felkner, L., McDonald, S.-K., Schneider, B., & Grogan, E. (2012). Female and male adolescents' subjective orientations to mathematics and the influence of those orientations on postsecondary majors. *Developmental Psychology, 48*(6), 1658–1673. https://doi.org/10.1037/a0027020

Robson, K., & Pevalin, D. (2016). *Multilevel modelling in plain language*. SAGE Publishing.

Rodríguez-Planas, N., & Nollenberger, N. (2018). Let the girls learn! It is not only about math … it's about gender social norms. *Economics of Education Review, 62*, 230–253. https://doi.org/10.1016/j.econedurev.2017.11.006

Smith, S. E., & Walker, W. J. (1988). Sex differences on New York state Regents examinations: Support for the differential course-taking hypothesis. *Journal for Research in Mathematics Education, 19*(1), 81–85. https://doi.org/10.5951/jresematheduc.19.1.0081

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE Publishing.

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25*(3), 246–280. https://doi.org/10.1080/08957347.2012.687650

Valero, P. (2017). Mathematics for all, economic growth, and the making of the citizen-worker. In T. S. Popkewitz, J. Diaz, & C. Kirchgasler (Eds.), *Political sociology and transnational educational studies: The styles of reason governing teaching, curriculum and teacher education* (pp. 117–132). Routledge.

Watt, H. M. G., Shapka, J. D., Morris, Z. A., Durik, A. M., Keating, D. P., & Eccles, J. S. (2012). Gendered motivational processes affecting high school mathematics participation, educational aspirations, and career plans: A comparison of samples from Australia, Canada, and the United States. *Developmental Psychology, 48*(6), 1594–1611. https://doi.org/10.1037/a0027838

Wheater, R., Ager, R., Burge, B., & Sizmur, J. (2013). *Student achievement in Northern Ireland: Results in mathematics, science, and reading among 15-year-olds from the OECD PISA 2012 study*. National Foundation for Educational Research. https://www.nfer.ac.uk/publications/pquk03/pquk03.pdf.

Woltman, H., Feldstain, A., Mackay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 52–69. https://doi.org/10.20982/tqmp.08.1.p052.

Zawistowska, A. (2017). Gender differences in high-stakes maths testing: Findings from Poland. *Studies in Logic, Grammar and Rhetoric, 50*(1), 205–226. https://doi.org/10.1515/slgr-2017-0025

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.