Supplementary methods – Mean Grade coding

Coding information about grade/school level was crucial due to the broad range of grades, from preschool to adulthood, covered in our meta-analysis. To account for this, we coded the grade for each sample. However, many samples spanned multiple grades, and in some cases, multiple school levels. As explained in the manuscript, we calculated the mean grade for each sample. When exact grade information was unavailable, we estimated it using the mean age (if provided) and/or details about the school level, aiming for the most accurate and unbiased approximation based on the available data. Grades were coded on a scale from 0 = kindergarten/preschool to 14 = university students or adults. Below is a list of the coding rules, with examples, that were applied to determine the mean grade. The list is organized in descending order of importance, meaning the first rules were prioritized, and subsequent rules were only applied when higher-priority rules could not be used.

Table S1

Coding rules with actual examples.

Coding rule	**Example**	
	Reported information	**Coded mean grade**
Directly reporting grade	3^{rd} grade children	3
Approximation based on median grade	2^{nd} to 5^{th} grade children	3.5
Approximation based on school level	primary school children	3
Approximating based on mean age	mean age of 10.5 years	5
Approximation based on age range	5 to 9-year-old age range	2.5
Grade or age range encompassing both children and adults	7 to 26-year-old age range	[not coded]

Supplementary results – Additional information on Moderator analysis

Table S2

Predicted sex differences (Cohen's d) across all four moderators from the full multilevel meta-regression.

Moderator	Level / Slope	Predicted effect	SE	95% CI
Type of content	Advanced Maths	-0.008	0.036	[-0.079, 0.064]
	Basic Numeracy	-0.002	0.043	[-0.085, 0.082]
	Broad Mathematics	0.145	0.026	[0.093, 0.197]
	Computation	0.089	0.033	[0.025, 0.154]
	Geometry	0.218	0.053	[0.115, 0.322]
	National Test	0.068	0.031	[0.007, 0.129]
Grade	Slope (β)	0.009	0.003	[0.005, 0.014]
Geographical area	Africa	0.026	0.068	[-0.107, 0.158]
	Central Europe	0.164	0.031	[0.103, 0.226]
	East Asia + Oceania	0.043	0.048	[-0.052, 0.138]
	East Europe + Russia	0.198	0.117	[-0.032, 0.428]
	Middle East	0.048	0.078	[-0.106, 0.202]
	North America	0.047	0.020	[0.007, 0.086]
	North Europe	0.003	0.049	[-0.092, 0.099]
	South + Central America	0.153	0.048	[0.059, 0.247]
GGGI	Slope (β)	0.921	0.471	[-0.003, 1.844]

Table S3

Pairwise contrasts between Geographical area levels based on estimated marginal means from the full multilevel meta-regression model. Values reflect unadjusted and Tukey-adjusted p-values. Positive estimates indicate a larger male advantage in the first region compared to the second.

Contrast	Estimate	SE	z	p (unadj)	p (adj)
Africa - CentralEurope	-0.139	0.077	-1.801	0.072	0.620
Africa - EastAsia&Oceania	-0.017	0.068	-0.245	0.807	1.000
Africa - EastEurope&Russia	-0.172	0.132	-1.304	0.192	0.898
Africa - MiddleEast	-0.022	0.079	-0.278	0.781	1.000
Africa - NorthAmerica	-0.021	0.071	-0.293	0.769	1.000
Africa - NorthEurope	0.023	0.096	0.234	0.815	1.000
Africa - SouthAmerica	-0.127	0.074	-1.714	0.087	0.678
CentralEurope - EastAsia&Oceania	0.122	0.059	2.081	0.037	0.427
CentralEurope - EastEurope&Russia	-0.034	0.121	-0.278	0.781	1.000
CentralEurope - MiddleEast	0.117	0.087	1.337	0.181	0.885
CentralEurope - NorthAmerica	0.118	0.034	3.451	0.001	0.013
CentralEurope - NorthEurope	0.161	0.053	3.058	0.002	0.046
CentralEurope - SouthAmerica	0.012	0.052	0.224	0.823	1.000
EastAsia&Oceania - EastEurope&Russia	-0.155	0.124	-1.253	0.210	0.916
EastAsia&Oceania - MiddleEast	-0.005	0.075	-0.071	0.944	1.000

EastAsia&Oceania - NorthAmerica	-0.004	0.052	-0.079	0.937	1.000
EastAsia&Oceania - NorthEurope	0.039	0.078	0.502	0.616	1.000
EastAsia&Oceania - SouthAmerica	-0.110	0.060	-1.836	0.066	0.595
EastEurope&Russia - MiddleEast	0.150	0.137	1.099	0.272	0.957
EastEurope&Russia - NorthAmerica	0.151	0.118	1.278	0.201	0.907
EastEurope&Russia - NorthEurope	0.195	0.128	1.516	0.130	0.799
EastEurope&Russia - SouthAmerica	0.045	0.125	0.364	0.716	1.000
MiddleEast - NorthAmerica	0.001	0.082	0.015	0.988	1.000
MiddleEast - NorthEurope	0.045	0.108	0.411	0.681	1.000
MiddleEast - SouthAmerica	-0.105	0.082	-1.286	0.198	0.904
NorthAmerica - NorthEurope	0.043	0.049	0.876	0.381	0.988
NorthAmerica - SouthAmerica	-0.106	0.051	-2.097	0.036	0.417
NorthEurope - SouthAmerica	-0.149	0.073	-2.034	0.042	0.459

Table S4

Pairwise contrasts between Type of content levels based on estimated marginal means from the full multilevel meta-regression model. Values reflect unadjusted and Tukey-adjusted p-values. Positive estimates indicate a larger male advantage in the first content area compared to the second.

Contrast	**Estimate**	**SE**	**z**	**p (unadj)**	**p (adj)**
AdvancedMaths - BasicNumeracy	-0.006	0.045	-0.133	0.894	1.000
AdvancedMaths - BroadMathematics	-0.153	0.033	-4.594	0.000	0.000
AdvancedMaths - Computation	-0.097	0.038	-2.547	0.011	0.111
AdvancedMaths - Geometry	-0.226	0.054	-4.198	0.000	0.000
AdvancedMaths - NationalTest	-0.076	0.032	-2.347	0.019	0.175
BasicNumeracy - BroadMathematics	-0.147	0.040	-3.667	0.000	0.003
BasicNumeracy - Computation	-0.091	0.038	-2.397	0.017	0.157
BasicNumeracy - Geometry	-0.220	0.059	-3.712	0.000	0.003
BasicNumeracy - NationalTest	-0.070	0.041	-1.688	0.091	0.539
BroadMathematics - Computation	0.056	0.030	1.848	0.065	0.434
BroadMathematics - Geometry	-0.073	0.051	-1.440	0.150	0.702
BroadMathematics - NationalTest	0.077	0.026	2.971	0.003	0.035
Computation - Geometry	-0.129	0.055	-2.349	0.019	0.175
Computation - NationalTest	0.021	0.032	0.650	0.515	0.987
Geometry - NationalTest	0.150	0.052	2.870	0.004	0.047

Supplementary results – Robustness check excluding multi-grade samples

An additional analysis was conducted as a robustness check to assess whether the use of aggregate statistics at the sample level biased the estimated moderator effect of Grade (many samples spanned multiple grades, and as no individual data were available, the average grade had to be coded as moderator in those cases). Specifically, we re-ran the meta-regression excluding all samples that spanned more than one grade level. This led to the removal of 24% of studies, 21% of samples, and 19% of effect sizes (resulting in a final $k = 955$). The model included only Grade as the moderator (i.e., no additional covariates for this robustness check). The estimated effect remained virtually unchanged ($B = 0.0066$, $SE = 0.0026$, $p = .013$), compared to the original estimate based on the full dataset ($B = 0.0065$, $SE = 0.0025$, $p = .010$). These results suggest that reliance on aggregate-grade coding, while suboptimal, did not meaningfully bias the observed association between grade and sex differences in mathematics.

Supplementary results – Metaforest analysis

To further investigate heterogeneity not accounted for by our primary moderator model, we conducted a random forest exploratory analysis using the "metaforest" package of R (van Lissa, 2020). This implements a machine learning approach based on random forests, trying to model non-linear relationships between moderators and effect sizes.

A limitation of metaforest is that it does not allow to account for the multilevel dependency structure present in our data. For this reason, we first fitted the full multilevel meta-regression model with all random effects, and then extracted the residuals. These residuals reflect the variation of the effect size that is not explained by the random effects (keeping into account the multilevel structure of dependencies). We used them as the outcome variable in the metaforest model.

The metaforest model included the same predictors coded for the structured moderator analysis: GGGI, Geographical Area, Grade, and Type of math content. In addition, we included other two pieces of information that were coded in the manuscript but had not been treated as moderators of interest in the main moderator analysis: Publication Year and Publication Type. We used 1,000 trees (double the model default value), with 2 candidate variables per split and a minimum terminal node size of 5 (the latter two values as per default in the metaforest package). The model achieved an out-of-bag (OOB) R^2 of 0.203, meaning that about 20% of the residual variance was explained.

The convergence plot (Figure S1 below) showed stable MSE reduction over the trees, supporting model adequacy. Partial dependence analysis is shown in Figure S2. This displays the marginal effects of each candidate moderator on residuals from the intercept-alone multilevel model (random effects only). Results suggest that Math content and Grade, and in part GGGI, are important moderators, consistently with the main moderator analysis, with effects tending to increase in later grades and varying across content domains. In contrast, other variables explain little of the remaining variance, at least in isolation.

Given the limitations of using residuals and the inability to model nested structures in MetaForest, these findings should be considered as exploratory. They nonetheless reinforce the conclusion that sex differences in mathematics are not uniform but depend on specific content domains and educational stages.

Figure S1

Convergence plot of the metaforest model, displaying cumulative Mean Squared Error (MSE) as a function of the number of trees.

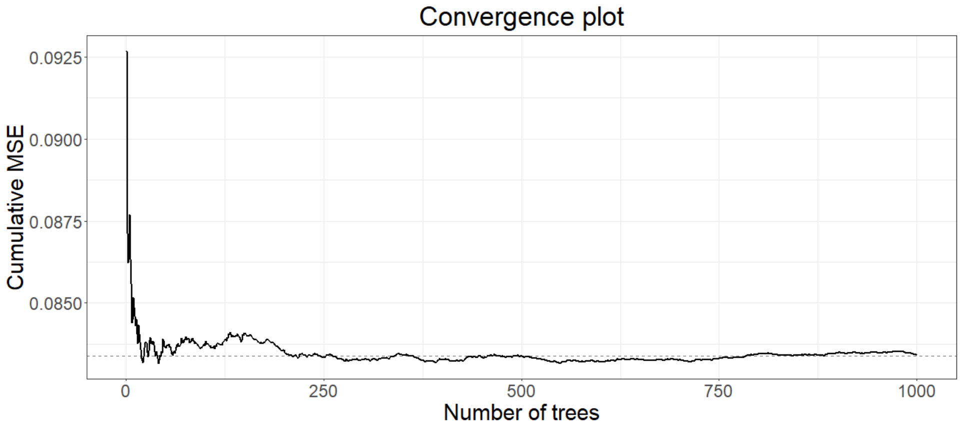
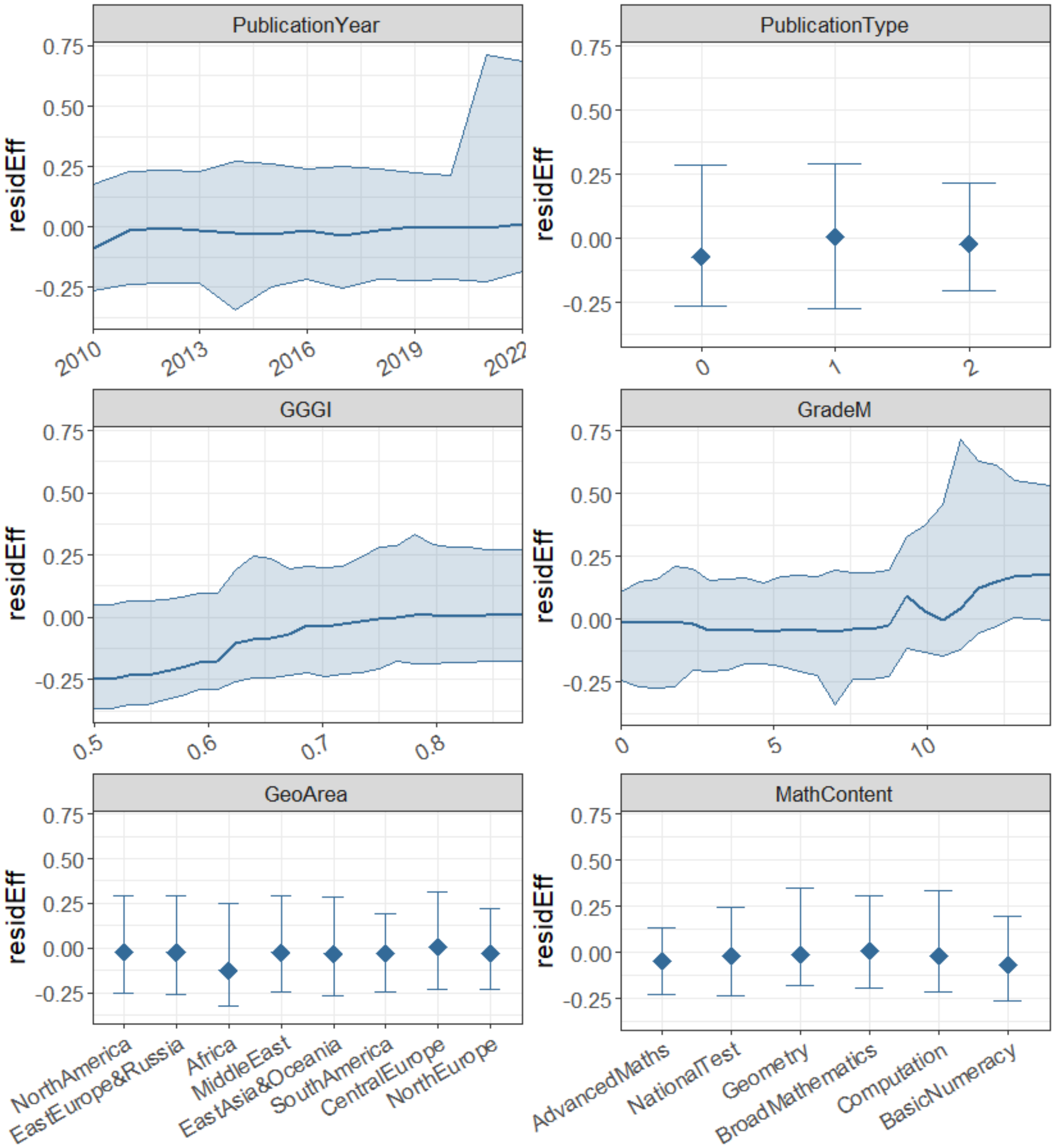


Figure S2

Partial dependence of residual effect sizes. "MathContent" is Type of math content.



Supplementary results – Additional information on Differential Item Functioning

To investigate the potential impact of Differential Item Functioning (DIF) on effect sizes, we specified an alternative multilevel meta-regression model. Although DIF is an important part of large-scale evaluations, only a small proportion of the studies in our sample explicitly reported whether DIF techniques were used, preventing coding of this moderator. To address this issue, we compared national and international large-scale evaluations, which often use some forms of DIF techniques (i.e., DIF Tests), to all other studies (i.e., No-DIF Tests).

The estimated effect size for DIF Tests showed sex difference advantaging males, $B_0 = 0.064$; $p = .001$, *95% CI* [0.0255; 0.1014]. The estimated coefficient, $B_1 = .028$, $p = .216$; *95% CI* [-0.0161; 0.0711], suggested sex difference even larger in No-DIF Tests (marginal mean: $d = 0.091$). Results suggested that including this moderator in our model did not improve the base model, in fact, the test for this moderator was not significant, $QM(df = 1) = 1.532$; $p = .216$. Conversely, the test for residual heterogeneity was still significant, $QE(df = 1173) = 582177.988$; $p < .0001$. The graph below shows the estimated marginal means of DIF tests contrasted with No-DIF Tests (Figure S3). These findings should be interpreted with caution; however, they support the hypothesis that correction for item bias may reduce to some extent sex differences in mathematical performance.

Figure S3

Estimated marginal means for of DIF tests contrasted with No-DIF Tests.