Uncovering Statistical Biases: A Bayesian Reanalysis of Reported Gender Disparities in STEM Education

Master Thesis

Name: Haaker, M.

Student number: s4357302

Master Ethics of Education: Philosophy, History, and Law

Faculty of Behavioural and Social Sciences

University of Groningen

Primary supervisor: Sarahanne M. Field

Second assessor: Dr. Pauline Schreuder

May 30th, 2025

Words: 7948

Summary

The thesis ‘Uncovering Statistical Biases: A Bayesian Reanalysis of Reported Gender Disparities in STEM Education’ assesses how prior findings on gender inequality in STEM education hold up under a Bayesian approach, using data from Guo et al. (2024). Replicability concerns raise questions not only about whether findings can be reproduced, but how they should be interpreted under different statistical frameworks. This study addresses that issue by asking: To what extent can a Bayesian reanalysis replicate and support the gender disparities reported by Guo et al. (2024)? The study focuses on country-level associations reported in the original article, comparing significance-based results to Bayes Factors via regression models. It also explores how methodological and epistemological choices inform interpretation in cross-national education research. The results showed partial alignment: some effects had a similar degree of evidence, while others yielded weaker or inconclusive Bayesian evidence, even when statistically significant under a frequentist framework. In other instances, Bayesian analysis indicated a higher degree of belief in an effect than the original study suggested. The study concludes that a reanalysis grounded in different methodological and epistemological underpinnings may support a contrasting lens through which to interpret findings. Moreover, it concludes that gender inequality in STEM education cannot be fully understood through country-level data alone, as such measures are often aggregated from a variety of indicators and carry sociopolitical assumptions about what constitutes equality, overlooking lived experiences. As such, the study calls for both methodological transparency and greater integration of contextual or qualitative approaches in future research.

Samenvatting

De scriptie ‘Uncovering Statistical Biases: A Bayesian Reanalysis of Reported Gender Disparities in STEM Education’ onderzoekt in hoeverre eerdere bevindingen over

genderongelijkheid in STEM-onderwijs standhouden onder een Bayesiaanse benadering, gebasseerd op data uit Guo et al. (2024). Zorgen over de reproduceerbaarheid van onderzoek roepen niet alleen vragen op of bevindingen herhaalbaar zijn, maar ook hoe ze geïnterpreteerd worden binnen verschillende statistische kaders. Deze studie behandelt die kwestie door de vraag te stellen: In hoeverre kan een Bayesiaanse heranalyse de gerapporteerde genderverschillen uit Guo et al. (2024) repliceren en ondersteunen? De studie onderzoekt landniveau-associaties uit het oorspronkelijke artikel en vergelijkt frequentistische resultaten met Bayes Factoren met regressiemodellen. Daarnaast wordt onderzocht hoe methodologische en epistemologische keuzes de interpretatie van bevindingen in cross-nationaal onderwijsonderzoek vormgeven. De resultaten lieten een gedeeltelijke overeenstemming zien: sommige effecten lieten een vergelijkbare mate bewijs zien, terwijl andere zwakkere of onduidelijke Bayesiaans bewijs opleverden, ondanks significante resultaten in de frequentistische analyse. In andere gevallen wees de Bayesiaanse analyse op een sterker vertrouwen in een effect dan het oorspronkelijke onderzoek suggereerde. De studie concludeert dat een heranalyse, gebaseerd op andere methodologische en epistemologische uitgangspunten, een ander perspectief op de interpretatie van bevindingen kan ondersteunen. Bovendien stelt de studie dat genderongelijkheid in STEM-onderwijs niet volledig begrepen kan worden op basis van kwantitatieve gegevens op landniveau, aangezien zulke maatstaven vaak zijn samengesteld uit verschillende indicatoren die sociopolitieke aannames bevatten over wat gelijkheid inhoudt en geleefde ervaringen over het hoofd zien. Daarom pleit de studie voor methodologische transparantie en meer integratie van contextuele of kwalitatieve benaderingen in toekomstig onderzoek.

Table of Contents

Introduction

In the recent past, a replication crisis was surging through the social sciences landscape (Field et al., 2019). Previous large-scale replication efforts revealed that only 36% of studies produce comparable results to the original findings (Open Science Collaboration, 2015). As a consequence, the credibility of pivotal discoveries has been obelized. Schnepf and Groeben (2024) argue that we have yet to fully overcome this crisis, but the first steps towards resolution have been taken (e.g., guidelines for more openness and transparency). However, these strategies are not effective enough because they are superficial and inherently preemptive (Field et al., 2019; Schnepf & Groeben, 2024). Thus, today's body of literature remains a jungle of uncertainty, where the chance of coming across reliable findings seems as haphazard as navigating one's way out of thick undergrowth. According to Field et al. (2019) and many others, the most direct way to resolve this issue is by replicating existing empirical research.

One area where the ground is ripe for gain is that of science, technology, engineering and mathematics (i.e., STEM) education; particularly in addressing persistent gender gaps. Although more women are entering higher education, they remain underrepresented in STEM (Spencer et al., 1999; Yazilitas et al., 2017). This underrepresentation extends beyond academia, as women make up less than 30% of all STEM workers (World Economic Forum, 2023). The lack of women in high-paying, high-skilled STEM jobs limits both workforce development and women's social mobility, impacting national and personal growth opportunities (Beroíza-Valenzuela & Salas-Guzmán, 2024). Moreover, this gender bias wastes valuable talent and potential and must be addressed to support fairer human development (Msambwa et al., 2024).

The underrepresentation of women is reinforced by the common stereotype that women are less capable in STEM fields (Beroíza-Valenzuela & Salas-Guzmán, 2024; Spencer

et al., 1999). This stereotype has long been falsely justified by the belief that this gendered skill gap is rooted in genetics (Spelke, 2005). For example, Klysing (2020) explains how psychological essentialism (i.e., the belief that social categories like gender reflect natural, inherent traits) often leads people to assume that gender differences in behavior or ability are biologically determined. Despite a lack of evidence for such a belief, consequences of these misconceptions have a widespread impact, as they negatively influence academic experiences and lead to discouragement in pursuing STEM-related degrees or professions for women (Bloodhart et al., 2020; Spencer et al., 1999). Moreover, the ways in which society influences boys and girls differently based on traditional gender expectations (i.e., gender-role socialization theory), results in unequal representation in STEM fields (Narh & Buzzelli, 2024; Spencer et al., 1999). For instance, boys are often encouraged to pursue STEM subjects while girls receive less support (Spencer et al., 1999). Similarly, students' self-concept and belief in their abilities play a crucial role in shaping behavior and are strongly linked to successful learning and performance (Dweck, 1999; OECD, 2015; Swafford & Anderson, 2020). These factors collectively contribute to the gender gap. Stereotypes shape the level of support someone receives, which in turn influences their self-belief and ultimately affects their performance.

Recent studies (e.g., Else-Quest, Hyde, & Linn, 2010; Stoet & Geary, 2018) have explored this issue with prominent implications, eliciting a need for replication in order to check robustness. The current study heeds this need by replicating and critically examining a recent cross-national study on gender disparities in STEM education. The first aim is to determine the reliability of the research by examining whether reported findings hold under a Bayesian reanalysis, or if methodological flaws and biases have affected conclusions. Rather than proving or disproving the existence of gender disparities in STEM, I focus on assessing whether a specific reported pattern holds up under alternative statistical scrutiny. In doing so,

I examine how different statistical approaches can provide distinct epistemological perspectives on evidence and confirmation. In a broader sense, my second aim is to show how the strength and interpretation of claims in cross-national educational research are framed by methodological choices.

The article by Guo et al. (2024) provides a proper foundation for reanalyzing of the reported statistical findings. The article focuses on three important STEM-related constructs: relative strength measures, aspirations, and graduation rates. They take a cross-cultural approach, relating those constructs to country-level gender equality measures. Their analysis showed a negative relationship between gender equality (i.e., GGGI) and relative science strength. This suggests that in countries that are more gender-equal, the gender gap in science achievement, relative to math and reading, was larger. However, this link disappears when education-specific equality measures (i.e., mean years of schooling, university enrollment, and expected years of schooling) were included. These measures were selected because they are directly linked to the three outcome variables (i.e., relative academic strength, STEM aspirations, and STEM graduation) and reflect early indicators of gender equality in education across societies. Based on this, they concluded that the observed differences are less about gender equality itself and more about broader systemic and cultural differences. This was also the case for participation in STEM (e.g., graduation rates) and STEM aspirations, which presented wide variability across countries, with no strong link to the gender equality measures. Moreover, STEM aspirations were found to be more related to general academic performance and relative math strength than relative science strength.

Bayesian epistemology is at the heart of my research when assessing the robustness of Guo et al.'s (2024) findings. Bayesian confirmation theory (BCT) is widely regarded as central to Bayesian epistemology (Strevens, 2017). According to Hawthorne (2011), BCT is a probabilistic approach to evaluating how evidence confirms or disconfirms scientific

hypotheses, using Bayes' Theorem to calculate the degree of confirmation. To illustrate how Bayesian epistemology works:

Suppose you believe there is a 70% chance it is going to rain today. That would be your prior belief (perhaps based on season or a weather forecast). Later, you look outside and see dark clouds forming. That is your new evidence, which makes you more confident it will rain. This leads you to update your belief to 90%.

Bayesian epistemology models belief updating in the same way: you start with a prior and then update your belief as new evidence becomes available. In this view, evidence is defined by the extent to which it shifts the degree of belief assigned to a hypothesis, not whether it crosses a threshold.

A premise of BCT is that all confirmation begins with some degree of prior plausibility. These priors influence how evidence is interpreted, but they are subjective: there is no single correct prior (meaning that two people can have different, reasonable starting beliefs). Indeed, using a default prior (such as a Cauchy) still carries assumptions about symmetry or scale. Moreover, within BCT, confirmation depends on the likelihood of observing the evidence given a hypothesis. However, these likelihoods are determined by auxiliary assumptions, background knowledge, and model structure, making them partly subjective. Thus, Bayesianism acknowledges the subjectivity of priors and model choices, while providing structured, transparent rules (Bayes' Theorem) for updating those beliefs. Bayes' Theorem is structured in a sense that it tells you how to update your beliefs based on new data, making the process internally consistent (i.e., everyone uses the same mathematical framework). It is transparent because it requires you to state your priors explicitly and explain why you chose them, as well as making it clear (through sensitivity analyses) how results change if priors change. Additionally, BCT emphasizes the theory-ladenness of confirmation: what counts as evidence depends in part on the theoretical presuppositions underlying the

model (Hawthorne, 2011; Strevens, 2017), especially in social science, where theoretical and cultural context is embedded in the data.

Thinking of evidence as probabilistic rather than all-or-nothing (i.e., accepting or rejecting the null hypothesis) makes Bayes' Theorem a useful tool for quantifying how strongly the data support one hypothesis over another. While this epistemology provides a conceptual foundation, Bayesian statistical inference is the methodological approach I take in reanalyzing the findings of Guo et al. (2024).

As large-scale research commonly influences policy and public discourse, another pertinent concept to consider is reproducibility. Reproducibility refers to the ability to draw the same results as another study using the same dataset and methods. This is often viewed as a way to assess the reliability of the original findings, although that view is contested within the metascience community. Given that Guo et al. (2024) synthesized data from multiple international sources using traditional statistical methods, it is important to assess whether their conclusions hold under alternative analytical approaches, such as Bayesian inference, primarily because cross-national data are inherently complex and combining them introduces heterogeneity. Reanalyzing the data using a different inferential paradigm serves as a targeted robustness check, indicating how sensitive the conclusions are to methodological assumptions.

With this context in mind, I seek to answer the following question: *To what extent can a Bayesian reanalysis replicate and support the gender disparities reported by Guo et al. (2024)?* To answer this research question, I consider two sub-questions:

(1) *How do the Bayesian reanalysis results compare to the original frequentist findings, and what methodological or epistemological challenges arise from this comparison?*

(2) *What are the implications of the reanalysis for understanding gender disparities in STEM education and how can Bayesian inference inform their interpretation?*

The reanalysis is performed under the assumption that if statistical biases have affected the findings of Guo et al. (2024), it will likely be reflected in discrepancies between the direction and strength of evidence between the reported frequentist tests and the Bayesian reanalysis.[1] In such cases, discrepancies between the findings may underscore the importance of replication. At the same time, this reanalysis may draw attention to how different methodological frameworks can lead to differences not only in results, but also in the interpretation of the same data. This reflection contributes to ongoing conversations about reproducibility of current research and emphasizes the importance of evaluating how evidence is produced and interpreted in studies addressing gender imbalance in STEM education.

My thesis will start with an overview of Guo et al.'s (2024) results. Next, I will provide an outline of the methodology, explaining the Bayesian reanalysis approach and its application to the used datasets. Following this, I will present a summary of my findings and compare Bayesian results to those of Guo et al. (2024). In the ensuing discussion and conclusion, I will discuss the two sub-questions and main research question and note the limitations of this study and directions for future studies.

G24 as a case

As has been made clear, gender disparities in STEM education remain a significant concern in both academic research and everyday practice. Although numerous studies have attempted to quantify these disparities, the reliability of their statistical interpretations

[1]It is also important to note that it is entirely possible to get statistical discrepancies between Bayesian and frequentist results that are not the result of bias but can arise from methodological differences such as priors or sample size effects.

warrants a closer look. This section situates the present study in relation to G24, giving an overview of their findings. I have listed all the results from G24 in Table 1.

Table 1

Results from G24

Row	Variables	Statistics	Significant
1	Gender gap in intraindividual science strength; GGGI	r = .39, 95% CI [.17, .57], p < .01	**Yes**
1.1	Gender gap in intraindividual science strength; GGGI	Linear vs. quadratic: F = .73, p = .394, Δdf = 1 Cubic model fits best: F = 5.03, p = .009, Δdf = 2; Δ11.1%, Adjusted R^2 = 9.03% Linear: β = −.17, SE = .21, 95% CI [−.59, .25] Quadratic: β = −.05, SE = .07, 95% CI [−.19, .08] Cubic: β = .16, SE = .05, 95% CI [.06, .27] All VIFs < 4	**Yes**
2	Gender gap in intraindividual math strength; GGGI	r = −.16, 95% CI [-.38, .07]	No
3	Gender gap in intraindividual reading strength; GGGI	r = −.17, 95% CI [-.38, .07]	No
4	Gender gap in intraindividual science strength; gender gap in university enrollment	r = −.01, 95% CI [-.24, .23]	No
5	Gender gap in intraindividual science strength; gender gap in mean years of schooling	r = −.07, 95% CI [-.29, .17]	No
6	Gender gap in intraindividual science strength; gender gap in expected years of schooling	r = .16, 95% CI [-.07, .37]	No

7	Gender gap in STEM aspirations; GGGI	r = .11, 95% CI [-.13, .34]	No
8	Gender gap in STEM aspirations; gender gap in expected years of schooling	r = .25, 95% CI [.03, .46]	**Yes**
9	Gender gap in STEM aspirations; gender gap in university enrollment	r = .16, 95% CI [-.08, .38]	No
10	Gender gap in STEM aspirations; gender gap in mean years of schooling	r = -.03, 95% CI [-.26, .20]	No
11	Gender gap in intraindividual science strength; GGGI	Linear vs. quadratic: F = 3.28, p = .075, Δdf = 1 Cubic > linear: F = 3.46, p = .038, Δdf = 2 Linear: β = .27, SE = .20, 95% CI [-.12, .67] Quadratic + cubic: β = -.21, SE = .08, 95% CI [-.37, -.04]; Cubic: β = .10, SE = .05, 95% CI [.00, .20]	**Yes**
12	Gender gap in intraindividual science strength; gender gap in university enrollment	r = .04, 95% CI [-.22, .30]	No
13	Gender gap in intraindividual science strength; gender gap in mean years of schooling	r = .07, 95% CI [-.18, .32]	No
14	Gender gap in intraindividual science strength; gender gap in expected years of schooling	r = .24, 95% CI [-.01, .46]	No
15	Gender gap in STEM aspirations; GGGI	r = .08, 95% CI [-.17, .33]	No

16	Gender gap in STEM aspirations; gender gap in expected years of schooling	r = .29, 95% CI [.04, .51]	**Yes**
17	Gender gap in STEM aspirations; gender gap in university enrollment	r = .13, 95% CI [-.14, .39]	No
18	Gender gap in STEM aspirations; gender gap in mean years of schooling	r = .17, 95% CI [-.08, .41]	No
19	GGGI; STEM graduation propensity	r = −.43, 95% CI [−.58, −.26]	**Yes**
20	GGGI; Actual % of women in STEM degrees	r = −.19, 95% CI [−.37, .01]	No
21	Mean years of schooling; STEM graduation propensity	r = −.35, 95% CI [−.51, −.16]	**Yes**
22	Mean years of schooling; Actual % of women in STEM degrees	r = .22, 95% CI [.03, .40]	**Yes**
23	Expected years of schooling; STEM graduation propensity	r = -.18, 95% CI [-.37, .02]	No
24	Expected years of schooling; Actual % of women in STEM	r = .36, 95% CI [.18, .52]	**Yes**
25	University enrolment; STEM graduation propensity	r = .09, 95% CI [-.11, .28]	No
26	University enrolment; Actual % of women in STEM	r = .02, 95% CI [−.18, .26]	No

27	STEM graduation propensity; Actual % of women in STEM; Gender gaps in aspirations; Gender gaps in science strength	*r* ranged from −.10 to .17	No

Note. Row 1–6 represent findings on gender differences in relative academic strength (PISA2018); Row 7–10 represent findings on gender differences in STEM aspirations (PISA2018); Row 11–14 represent findings on gender differences in relative academic strength (PISA2015); Row 15–18 represent findings on gender differences in STEM aspirations (PISA2015); Row 19–27 represent findings on gender differences in STEM graduation (PISA2018). GGGI = Global Gender Gap Index.

Methodology

To evaluate Guo et al.'s (2024) results, I will perform a Bayesian reanalysis of the data to investigate alternative interpretations and assess the reliability of reported gender disparities in STEM education. Generally, I hypothesize that statistical biases (which are known to be pervasive in the behavioral sciences literature) in the original analysis will result in inaccurate estimates of gender disparity in STEM education. Previous Bayesian reanalyzes (see e.g., Field et al., 2019) have demonstrated that it is unlikely that all tests conducted in the original article can be sufficiently reproduced even when the same data are the subject of the analysis. *P*-values can either be shown to be too modest for the strength of the effects observed in in the data, or to overestimate the effect. The study hypothesis, therefore, is kept broad so as to enable flexibility and inclusivity in the analysis and avoid any preconceived bias. Thus, this study systematically analyses secondary data, executed through JASP and R software, using a Bayesian approach, rather than a frequentist approach.

The reason for choosing a Bayesian analysis is two-fold. First, is its ability to provide relative evidence both for and against the null hypothesis by assessing how strongly the data support it compared to the alternative, which helps address false positives (i.e., Type I errors) and false negatives (i.e., Type II errors). It is important to know if these types of inaccuracies occurred, as they may lead to incorrect conclusions. In the first case, the data might actually support the null hypothesis even if the original findings show a significant p-value. A Bayesian analysis can show stronger pro-null evidence in this case. In the second case, the original study may report non-significant p-values, even though there is a real effect. Here, a Bayesian analysis can give evidence in favor of the alternative hypothesis by comparing how strongly the data support it relative to the null. Unlike frequentist statistics, which only allow a binary decision threshold of 'reject the null' or 'fail to reject the null', Bayesian methods present a more nuanced view in both cases by estimating how likely the data are under

different hypotheses. It shows whether the evidence favors the presence of an effect or not given the data.

Second, the Bayesian approach can also be valuable to indicate when more data are needed. Under the frequentist framework, low power (insufficient data) leads to failing to reject the null, but gives no further information and can incorrectly cause researchers to assume that no effect exists (an informal fallacy: absence of evidence of an effect is not evidence of the absence of an effect, in this case). Being able to test for the presence *or* absence of an effect, *or* see if more data should be collected, is powerful.

As mentioned, a previously selected paper will be central in this study: *'Cross-Cultural Patterns of Gender Differences in STEM: Gender Stratification, Gender Equality and Gender-Equality Paradoxes'* by Guo et al. (2024). From here on, the article will be referred to as: 'G24'. I identified the article through a literature search in the ERIC database using the following search terms: "STEM education", "gender differences" or "gender inequality", and "international". The inclusion criteria for the search required that the paper be peer-reviewed and published within the timeframe of 2019 to 2024. Additionally, I selected the article due to its comprehensive nature and methodologically rigorous analysis. G24 did not just include relative science test scores in their analysis but incorporated other factors such as aspirations and graduation rates. Moreover, they established a well-thought-out methodology, making use of large-scale, high-quality data, constructing specific variables instead of using raw scores, and applying advanced statistical models (e.g., multilevel models). During the analysis, they took a novel approach to measuring students' relative science strength by comparing them to students' (relative) math and reading strengths, as opposed to just looking at science scores alone. Taking it even further, they connected the gender difference in science strength to national-level gender equality. Combining individual-level data with country-level context, offers an integrated view of both individual and

contextual factors. Collectively, these aspects show how the article presents a strong empirical foundation for reanalysis.

G24 mainly made use of two PISA datasets: PISA2015 and PISA2018. PISA typically evaluates the science, math, and reading performance of 15-year-olds using nationally representative samples (OECD, 2019). According to G24, PISA2015 includes 416,690 adolescents from 62 countries/regions. PISA2018 includes 528,681 adolescents from 71 countries/regions. Notably, the reason for analyzing PISA2015 was to replicate the PISA2018 analyses to see whether there were significant differences between the two. In addition to the PISA database, they incorporated an OECD dataset on the number of graduates by field of education (i.e., STEM) and several national indicators of gender equality retrieved from UNESCO: GGGI, the attendance rate of tertiary education, mean years of schooling, and expected years of schooling. For the OECD and the UNESCO datasets, they used the years 2015 and 2018 in order to match them up with the PISA datasets.[2]

In the current study, I will focus solely on the country-level research questions from G24 (i.e., RQ 1, 2 & 4)[3], as these findings are often used to make broad claims with wide implications. A Bayesian reanalysis is particularly suited to this level of analysis because it allows for a more detailed assessment of the strength of evidence, while acknowledging the theory-ladenness of data interpretation in the social sciences. On top of that, it narrows the scope, making the reanalysis more focused and manageable.

I will use the same methodological approach as G24 to ensure consistency and comparability. In the following sections, an explanation of how the measures were computed in the original study (and thus in mine) will be demonstrated, as this is essential for

[2]In the current research, countries with scores that were NA for one or both variables in a regression analysis were taken out.
[3]RQ1: How are gender differences in relative science strength (science achievement compared to math and reading) associated with country-level gender equality?
RQ2: How are gender differences in STEM career aspirations associated with country-level gender equality?
RQ4: How are gender differences in STEM graduation associated with country-level gender equality?

understanding the analysis. In addition, I will clarify the method of analysis employed in G24's study, which will also be adopted in this reanalysis.

Gender Differences in Relative Science Strength

The dependent variables include students' relative academic strength (science, math, and reading achievement), STEM-related career aspirations, and STEM graduation.

With regards to student academic performance, PISA estimated and reported 10 plausible values for each subject to represent student performance in their dataset. As all 10 plausible values were included in G24's analysis, they will also all be used in the current study. To calculate students' intraindividual academic strength, I followed the procedure outlined by Stoet and Geary (2018), as adopted by G24. (1) First, science, math, and reading achievement were standardized within each country to produce z-scores for each subject. (2) Next, a standardized average score of the three new z-scores was calculated for each student (zGeneral). (3) Lastly, each student's intraindividual science strength score was derived by subtracting zGeneral from zSience, and then that score was standardized again. To measure the national gender gap in science strength, I calculated the average science strength for boys and girls in each country. Then, I subtracted the girls' average from the boys' average. The national gender gap in intraindividual science strength was structured so that high values indicate high gender inequality, with boys outperforming girls in relative science scores.

Gender Differences in STEM Aspirations

Adhering to G24's method, I coded STEM aspirations as a binary variable (1 = STEM; 0 = non-STEM) following PISA's coding strategy (OECD, 2016, p. 283). I calculated the national gender gap in STEM aspirations by dividing the percentage of female students who aspire STEM careers by the percentage of male students who did so. As in G24, scores greater than one portray higher STEM aspirations among girls, whereas scores less than one portray higher aspirations among boys.

Gender Differences in STEM Graduation

In G24, gender differences in STEM graduation was reviewed by using two indicators: the propensity of women to graduate with STEM degrees and the actual percentage of women among people who earn STEM degrees.

I derived the first measure using the formula reported in G24: $a / (a + b)$, where a is the percentage of women who graduate with STEM degrees, and b is the percentage of men who graduate with STEM degrees. This measure compares how likely women are to graduate in STEM relative to men, independent of the gender differences in the overall number of graduates. The second measure reflects the share of women who acquired STEM degrees within the total population of STEM graduates during a given timeframe (in this case: 2018).

Country-Level Gender Equality Measures

As in G24, I used the Global Gender Gap Index (GGGI) as a composite measure of national gender equality. It includes 14 indicators (e.g., income, life expectancy) to assess gender differences within countries.

On top of that, G24 used three relative difference measures of national gender equality in the education domain: the attendance rate of tertiary education, mean years of schooling, and expected years of schooling. Enrollment rate of tertiary education denotes "the ratio of total enrollment (in tertiary education), regardless of age, to the population of the age group that officially corresponds to the level of education shown" (Guo et al., 2024, p. 36). Mean years of schooling refers to "the average number of years of education received by people ages 25 or older, converted from education attainment levels using official durations of each level." (Guo et al., 2024, p. 37). Expected years of schooling (which refers to the 'school life expectancy' dataset from UNESCO) means "the number of years of schooling in a country that children of school entrance age can expect to receive if prevailing patterns of age-specific enrolment rates persist throughout the child's life." (Guo et al., 2024, p. 37). A within-country

relative approach was used to compute gender equality indices, by taking the ratio female to male values (e.g., mean years of schooling for women divided by that for men).

Finally, as my thesis focuses on country-level analyses, individual-level control variables such as student gender and family socioeconomic status (SES), included in G24's multilevel models, were not applied here.

Data Analysis

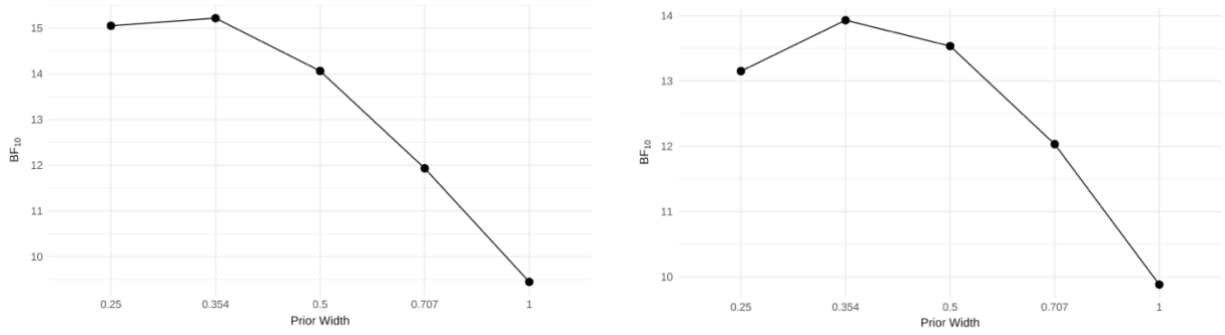
For the data analysis, I will apply the same approach in my study, that G24 explains. Of course, I will conduct the analyses through a Bayesian framework, as opposed to the frequentist used in G24. Moreover, G24 conducts all models with Mplus 8.1, whereas in this study I will use JASP and R. I will employ JASP for all the analyses in this study, whereas R can be considered a supplemental tool in constructing the datasets. I will apply Bayesian linear regression for the country-level analyses examining the relationship between gender equality and gender gaps in STEM indicators. I conducted all the analyses using JASP's default JZS prior (location = 0, scale = .354). I chose this prior because it is commonly used for Bayesian hypothesis testing and fits the type of analysis I am doing: it starts with the assumption that small effects are more likely, but it still allows the data to support larger effects if present.[4] I performed two sensitivity analyses by varying the prior width (r = 0.25, 0.354, 0.5, 0.707, and 1.0) for the associations between gender gap in relative science strength and GGGI, and STEM graduation propensity and GGGI (Figure 1). The results were consistent across prior widths, indicating that the evidence in favor of an effect remains robust across different prior assumptions. It is worth noting that the reanalyzes were initially conducted within a frequentist framework to ensure consistency with G24's results. This step

[4]This prior is based on a Cauchy distribution, which assigns some probability to all possible effect sizes, no matter how large.

was essential for comparability, confirming that any differences observed in the Bayesian analyses were accurate and valid.

Figure 1

Sensitivity Analyses Results



Note. Sensitivity analyses of the Bayes Factor (BF_{10}) across prior widths. The left panel displays results for the relationship between gender gap in relative science strength and GGGI. The right panel displays results for the relationship between STEM graduation propensity and GGGI.

Results

In this section, I will report the results from the Bayesian reanalysis and compare them to the results from G24. Rather than presenting each model result individually, I will focus on summarizing patterns across the tested models. The results are organized by outcome type: relative academic strength, STEM aspirations, and STEM graduation. I included a visual overview to illustrate the range and distribution of findings (Figure 2) and a more detailed table of the Bayesian results can be found in the Appendix.

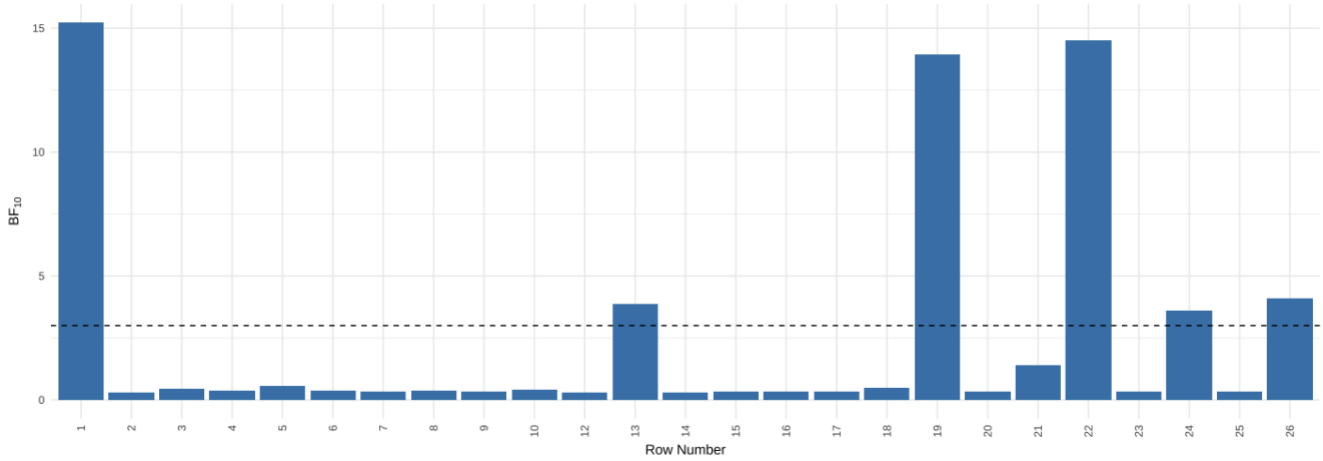
Overview of Bayes Factors

A total number of 25 models were tested and their Bayes Factors are summarized in Figure 2. The majority of models showed a Bayes Factor below the frequently used threshold of 3,

indicating weak or inconclusive evidence (Dittrich, Leenders, & Mulder, 2017).[5] Then there are three that float around that line (i.e., row 13, 24, and 26), making them borderline cases. The evidence is moderate and possibly leaning toward support for an effect. Only a small number demonstrated strong evidence, suggesting that a few predictors may play a more substantial role.

Figure 2

Visual Representation of Bayesian Results



Note. Each bar represents the BF_{10} for a specific model-row combination. The row numbers on the x-axis correspond to those in the Appendix. The dashed line at $BF_{10} = 3$ indicates the common threshold for moderate evidence in favor of an effect. Rows 1.1 and 11 are excluded because they involve model comparisons. Row 27 is excluded because it contains multiple BF_{10}s (ranging from .369–.512).

Relative Academic Strength

Across 9 tested models, only the relationship between GGGI and gender gap in relative science strength in PISA2018 showed strong evidence in favor of an effect, with a positive direction ($BF_{10} = 15.22$, $M = .661$). Coinciding with G24's significant result, it indicates that the data are about 15 times more likely under the alternative model than under

[5]It is good to note that this Bayesian 'threshold' does not involve the same reliance on strict cut-off points as frequentist statistics; it serves as a heuristic guideline, not a hard rule.

the null model, and that the gender gap was larger in more gender-equal countries. In contrast to science, the results for math and reading, as well as the three gender equality indices, were too weak to support a meaningful interpretation. Concerning PISA2015, the relationship between mean years of schooling and gender gap in relative science strength showed moderate support for an effect ($BF_{10} = 3.859$). Although G24 did not find this relationship statistically significant, the present results indicate that mean years of schooling may be associated with gender differences in academic science strength, though this warrants further investigation. The results regarding the other two gender equality indices were inconclusive.

For the relationship between GGGI and the gender gap in relative science strength, the cubic model was the most supported for this sample (P(M|data) = .450). Including the cubic term improved model fit, as reflected in a high inclusion Bayes Factor ($BF_{inclusion} = 41.457$), while the overall model showed moderate-to-strong support compared to all other models ($BF_M = 9.011$).[6] Regarding the PISA2015 findings for the same relationship, the best-fitting model included all three terms (linear, quadratic, and cubic) with the highest posterior probability (P(M|data) = .516), although evidence in favor of this model against all other models was moderate ($BF_M = 3.195$). This indicates that the non-linear relationship is plausible within this sample, but not strongly supported. Also, the quadratic term improved model fit (again, within this sample) as indicated by a high inclusion Bayes Factor ($BF_{inclusion} = 14.396$), which differs slightly from G24 where both quadratic and cubic terms were significant. Although there was moderate evidence suggesting that including a cubic term also improves model fit ($BF_{inclusion} = 4.360$), its contribution appears more limited than suggested by G24. Given the complexity of cross-national data, these patterns may reflect overfitting rather than a robust or interpretable effect. Further testing in new datasets would be required to assess overall strengths and generalizability.

[6] I standardized GGGI beforehand to avoid multicollinearity.

STEM Aspirations

Between the 8 models regarding STEM aspirations that were tested across PISA2015 and PISA2018, all came out inconclusive, meaning that the data are barely more likely under one model than the other. Although this indicates that there is not enough evidence to support a strong conclusion, G24 did find a significant relationship between gender gap in expected years of schooling and gender gap in STEM aspirations for both PISA2015 and PISA2018. These discrepancies, between the Bayesian inconclusive findings and the frequentist significant findings, may have occurred due to differences in sample composition or analytical approach.[7]

STEM Graduation

Among the models predicting female STEM graduation propensity and actual percentage of women in STEM, only two showed strong evidence for an effect: GGGI and female STEM graduation propensity (BF_{10} = 13.931), and mean years of schooling and actual percentage of women in STEM (BF_{10} = 14.502). For the first, the direction of the effect was negative (M = -.38), suggesting countries with higher gender equality tend to have lower female STEM graduation propensity. The direction of the latter was positive (M = .63), indicating that countries where women have more years of schooling relative to men tend to have a higher percentage of women in STEM fields. These relationships should be interpreted cautiously, though, given that broader systemic differences not captured by the current model may play a role. Again, generalizability is limited, and these findings should be investigated further in future research.

Additionally, there was some support for a positive association between actual percentage of women in STEM and expected years of schooling (BF_{10} = 3.615, M = .09). In comparison to G24, who found a significant relationship, the evidence here is moderate.

[7]This discrepancy will be investigated further in the discussion.

Contrastingly, they found no significant relationship between actual percentage of women in STEM and university enrolment, when again the current evidence is moderate ($BF_{10} = 4.111$, $M = .09$). While the evidence is moderate (suggesting the effect might be real), the effect sizes are small and may be sensitive to sample or model changes. Careful interpretation and future research are therefore necessary, as the practical significance of these findings is limited. Another interesting discrepancy happened between G24's significant finding for the link between mean years of schooling and STEM graduation propensity, and the inconclusive Bayesian results ($BF_{10} = 1.400$). This suggests that the evidence for this relationship is not strong enough to clearly support or reject the presence of an effect and may not be generalizable across samples or statistical approaches.

Discussion

In this thesis, I set out to determine the reliability of prior research on gender disparities in STEM education by examining reported findings using a Bayesian approach. The main research question was: *To what extent can a Bayesian reanalysis replicate and support the gender disparities reported by Guo et al. (2024)?* This question was then split into two sub-questions: (1) *How do the Bayesian reanalysis results compare to the original frequentist findings, and what methodological or epistemological challenges arise from this comparison?* (2) *What are the implications of the reanalysis for understanding gender disparities in STEM education and how can Bayesian inference inform their interpretation?* To address each sub-question, I structured the discussion to examine them both individually, with the first sub-question linked to methodological insights, and the second to substantive implications for gender in STEM. Following this, I will discuss limitations and suggestions for future research.

Methodological Insights

The Bayesian reanalysis showed partial alignment with G24's findings, with some predictors yielding similar effects, but most diverging in strength or certainty. For example, the relationship between gender gap in relative science strength and GGGI had a large Bayes Factor, aligning with the original frequentist finding of a significant *p*-value ($p < .01$). From a frequentist perspective this would mean that it is unlikely that the observed relationship occurred by chance under the null hypothesis. In Bayesian terms, a large Bayes Factor means the data provide strong evidence for the presence of an effect, relative to the null hypothesis. Thus, in this case both approaches suggest an effect, but the Bayesian analysis adds epistemological value by giving a more informative and interpretable understanding of the evidence by showing how strong the evidence is (i.e., treating evidence as a matter of degree, not a yes or no).

Contrastingly, the association between actual percentage of women in STEM and expected years of schooling was deemed significant under a frequentist framework. However, in comparison this translated into only moderate Bayesian evidence. This is not just a difference in result, but in what is fundamentally defined as confirmation: looking through a frequentist lens can potentially overstate the strength of evidence because it only suggests how rare the data would be if the null were true, not how plausible the hypothesis itself is. By contrast, the Bayesian result tempers that confidence (in evidence) by showing that the evidence only moderately supports the hypothesis but is not compelling. In epistemological terms, it essentially indicates how your degree of belief should change based on the data, relative to your prior. Regarding this case, this implies that one's belief in the hypothesis should increase, but only modestly.

As explained earlier, the interpretation of evidence depends not only on the data, but also on priors and model assumptions, which are inherently subjective. While I addressed the

potential influence of subjective prior choices through a sensitivity analysis, (i.e., see Figure 1), this does not eliminate the underlying epistemological issue of confirmation not being purely objective. However, Bayesianism does not view this subjectivity as a flaw, but as an integral part of scientific reasoning offering a framework (Bayes Theorem) for incorporating it. For instance, the JZS prior assumes that small to moderate effect sizes are more plausible than large ones, making it harder to provide evidence in favor of an effect. Despite that, the Bayesian analysis concerning the relationship between actual percentage of women in STEM and university enrolment indicated a moderate Bayes Factor, while the frequentist approach found it insignificant. As the prior was conservative, this moderate evidence for an effect increases confidence in the result (especially when compared to the frequentist finding), because the result does not depend on optimistic assumptions. More importantly, by explicitly stating and testing the influence of my prior, I made a subjective element of the analysis transparent, which is something frequentist statistics do not do. This illustrates the epistemological strength of Bayesian analysis in handling subjectivity.

In social science, hypotheses are very context-dependent, often involve values, and are harder to separate from, for example, cultural influences. Consequently, belief updating becomes more complex because data are ingrained in their own contexts, assumptions are often derived from theory, and variables (e.g., university enrolment, GGGI) are not objective. Although Bayesian inference does not remove this theory-ladenness, it makes such assumptions explicit through the specification of priors and shows, through transparent belief updating, how theoretical assumptions influences the interpretation of evidence. For instance, the GGGI is not a raw measure of gender inequality but a composite index based on 14 indicators, each based on presupposed ideas about which gaps matter and how they should be quantified.[8] One assumption behind the GGGI is that equal access to education is important

[8]G24 takes this fact into account in their study by including the analyses with three education-related gender equality measures. I will discuss this point further in the next section.

for gender equality. A high GGGI score would suggest that boys and girls have more equal educational opportunities. In this case, one might use a prior that assumes gender gaps are smaller in those countries with a higher GGGI score. The belief updating process then shows how much the data challenge or support that belief, making the reasoning behind the interpretation of the results explicit.

Substantive implications for gender in STEM

Although Bayesian inference does not explain why gender disparities occur, it may be a useful tool as it can indicate how strongly the data support their existence within a given sample. For instance, in my reanalysis the Bayes Factor suggested strong evidence for a relationship between gender equality (GGGI) and female STEM graduation propensity (in the PISA2018 sample). Other studies could replicate this by applying Bayesian inference to other PISA cycles to assess whether this is a consistent pattern across time or in particular countries. Based on those findings, researchers could then explore the 'why?' through contextual or qualitative research. Unlike frequentist analysis, which may treat non-significant results as inconclusive or dismissive, Bayesian inference could provide a clearer sense of how much support the data offer for or against a relationship. In the previous example, the strong Bayesian evidence suggests that the relationship is meaningful (within the sample), not just statistically detectable. On the other hand, in cases where only weak or moderate evidence is found, Bayesian inference helps avoid overinterpreting null results by distinguishing between absence of an effect and insufficient data. Still, interpreting what this evidence means requires more than just statistical tools.

This is especially important in studies like the present one, where gender inequality is analyzed at the country level and findings may reflect complex social realities as opposed to clear causal patterns. The results suggest a need for more critical and context-sensitive interpretations, rather than reinforcing firm conclusions about gender inequality in STEM,

particularly when based on country-level data. For instance, a higher gender equality score may coexist with entrenched gender stereotypes in a country. This can make it seem like equality exists, when in reality girls are still discouraged and underrepresented in domains like STEM. Indeed, even if patterns are observed, they can be hard to interpret because national indicators hide differences between and within countries.

This raises the question of whether national-level indicators are even appropriate for the kinds of gendered experiences being examined. Gender inequality is a concept that varies across contexts, yet aggregated indicators are based on 'objective' assumptions about what matters across all contexts. However, such assumptions are anything but neutral: they reflect sociopolitical values about what counts as equality, whose experiences are prioritized, and which forms of inequality are visible or measurable. For example, using the 'university enrolment' measure means that what counts as educational gender equality is formal participation, experiences of those who reach higher education are prioritized, and what is made visible is institutional access, not cultural norms or microaggressions. Instead of seeking universal conclusions about gender inequality in STEM, the findings hint at the importance of examining how such disparities crystallize within specific cultural, institutional, and educational contexts.

Reflecting on the limits of statistical reanalysis itself is therefore important. Gender inequality is not a natural law like gravity, it is a socially constructed and lived experience. Reducing it to statistical relationships or aggregated indicators may oversimplify the experiences of real people and sideline the voices of those most affected. In some cases, it may even be ethically problematic to treat gender inequality as a quantifiable phenomenon, as it passes over the stories and struggles behind the numbers. Ultimately, these findings suggest that interpreting gender disparities in STEM education requires not only statistical insight and

awareness of how gender is constructed and measured across various contexts, but also attention to the real-world experiences behind the data.

Limitations and suggestions for future research

There were several limitations to this study, especially related to methodological and data-related challenges that may explain some of the deviations from G24's findings. One issue involved retrieving the original datasets. Although G24 flawlessly documented the sources for PISA and the gender equality indices, the datasets on graduation rates were more difficult to locate. In particular, the dataset used for the actual percentage of women in STEM degrees could not be found in UNESCO's records. After careful consideration, I assumed it came from the OECD (since they are closely related to UNESCO), but G24 did not specify this.

Further challenges emerged during the computation of certain variables. For instance, G24 does not mention whether sampling weights were used in calculating relative science strength, nor do they specify how the plausible values (PVs) were handled during standardization.[9] Additionally, they omit a clear description of how the final relative science strength score was standardized. Stoet & Geary (2018) do mention this in their method section, although briefly.

There were also inconsistencies in the number of countries included in G24's analysis. While they claim to use 71 countries for PISA2018 (even though the full dataset included 72 countries), their supplementary file list only 61. Similar discrepancies appear for PISA2016, where they report 62 countries, but their (supplementary) table include 66, while the full dataset lists 73. In both cases, it remains unclear which countries were excluded and why.

[9] i.e., starting with 10 PVs, taking the average of those 10 PVs, and then standardizing that average score; or starting with 10 PVs, standardizing all 10 PVs, and then taking the average of those 10 standardized PVs.

Furthermore, mismatches between reported country scores and those found in the public datasets raise questions about data accuracy and potential errors in the original study.

In short, a lack of transparency around data selection, transformations, and variable construction in G24 made exact replication difficult. However, publicly available data is prone to get updated once in a while, which may explain some differences in the datasets. The methodological and documentation limitations identified in this study should also be an incentive for researchers to fully explain how the data were prepared and handled, particularly for large-scale, publicly available data.

Future research should focus on replicating the individual-level analyses from G24 using a Bayesian framework to evaluate their soundness. Such replication would be valuable for appraising whether broader national trends observed in G24 are supported or contradicted by micro-level patterns. As in G24 and related studies, exploring cross-cycle replication remains a compelling approach to evaluate temporal stability of observed relationships. Identifying whether such patterns are constant across time or tied to specific cycles can enhance (or decrease) confidence in their generalizability. At the same time, the nuanced and often inconclusive findings of this reanalysis emphasize the need for alternative methodological strategies that allow for a more context-sensitive interpretation. Future studies should carefully consider whether national-level data are suitable for capturing the complex, lived experiences of gender inequality, and keep an eye on the sociopolitical assumptions underlying the indicators used to represent it. While quantitative research can be valuable in this case, it may benefit from being complemented by qualitative or contextual research.

Conclusion

In summary, I used a Bayesian reanalysis to revisit the reported gender disparities by Guo et al. (2024). By reviewing a selection of country-level research questions, the first aim was to test whether the study could be replicated and if the original findings would hold under

this alternative method. Building on that, my second aim was to illustrate how methodological choices frame the strength and interpretation of findings in cross-national educational studies.

The results demonstrated partial alignment with G24: some effects were similar, but most were inconclusive, demonstrating how research conclusions can be sensitive to the methods used to evaluate evidence. The Bayesian method provided insight into the degree of support for both presence and absence of effects, indicating weak or inconclusive evidence where G24 reported significance, and in some cases, suggested stronger support for an effect where the original analysis had not.

An overarching theme in this study is the importance of methodological transparency and caution with interpreting cross-national data on gender inequality, as the tools used to measure it are filled with assumptions and may not fully capture lived realities. This reanalysis suggests that alternative statistical approaches can provide a different lens through which to interpret complex phenomena and potentially support more reflective and transparent interpretations in educational research.

Ultimately, the value of this reanalysis lies not in overturning previous findings, but in reminding us that even widely accepted notions or methods deserve a second look; not to undermine their previous contributions, but to rethink the confidence we place in them.

References

Beroíza-Valenzuela, F., & Salas-Guzmán, N. (2024). STEM and gender gap: A systematic review in WoS, Scopus, and ERIC databases (2012–2022). *Frontiers in Education*, *9*, Article 1378640. https://doi.org/10.3389/feduc.2024.1378640

Bloodhart, B., Balgopal, M. M., Casper, A. M. A., Sample McMeeking, L. B., & Fischer, E. V. (2020). Outperforming yet undervalued: Undergraduate women in STEM. *PLOS ONE*, *15*(6), e0234685. https://doi.org/10.1371/journal.

Dittrich, D., Leenders, R. T. A. J., & Mulder, J. (2017). Network autocorrelation modeling: A

Bayes factor approach for testing (multiple) precise and interval hypotheses. *Sociological Methods & Research*, *46*(4), 784–815. https://doi.org/10.1177/00491

Dweck, C. S. (1999). *Self theories: Their role in motivation, personality, and development*. (1st ed.). Psychology Press. https://doi.org/10.4324/9781315783048

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

Field, S. M., Hoekstra, R., Bringmann, L., & van Ravenzwaaij, D. (2019). When and why to replicate: As easy as 1, 2, 3? *Collabra: Psychology, 5*(1), Article 46. https://doi.org/10.1525/collabra.218

Guo, J., Marsh, H. W., Parker, P. D., & Hu, X. (2024). Cross-cultural patterns of gender differences in STEM: Gender stratification, gender equality, and gender-equality paradoxes. *Educational Psychology Review, 36*(2), 1–48. https://doi.org/10.1007

Hawthorne, J. (2011). Bayesian confirmation theory. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Philosophy of statistics* (pp. 17–100). Elsevier. https://doi.org/10.1016/B978-0-444-51862-0.50010-1

Klysing, A. (2020). Exposure to scientific explanations for gender differences influences individuals' personal theories of gender and their evaluations of a discriminatory situation. *Sex Roles, 82*(5–6), 253–265. https://doi.org/10.1007/s11199-019-01060-w

Msambwa, M. M., Daniel, K., Lianyu, C., & Antony, F. (2024). A systematic review using feminist perspectives on the factors affecting girls' participation in STEM subjects. *Science & Education*. https://doi.org/10.1007/s11191-024-00524-0

Narh, E. D., & Buzzelli, M. (2024). Women on the move for science, technology, engineering and mathematics: Gender selectivity in higher education student migration. *Higher Education Quarterly*, *78*(3), 745-765. https://doi.org/10.1111/hequ.12483

OECD. (2015). *The ABC of gender equality in education: Aptitude, behaviour, confidence* (PISA). OECD Publishing. https://doi.org/10.1787/9789264229945-en

OECD. (2016). *PISA 2015 results (Volume I): Excellence and equity in education*. OECD Publishing. https://doi.org/10.1787/9789264266490-en

OECD. (2019). *PISA 2018 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Schnepf, J., & Groeben, N. (2024). The replication crisis as mere indicator of two fundamental misalignments: Methodological confirmation bias in hypothesis testing and anthropological oversimplification in theory-building. *New Ideas in Psychology, 75*, 101110. https://doi.org/10.1016/j.newideapsych.2024.101110

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science?: A critical review. *American Psychologist*, *60*(9), 950-958. https://doi.org/10.1037/0003-066X.60.9.950

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4-28. https://doi.org/10.1006/jesp.1998.1373

Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science, 29*(4), 581–593. https://doi.org/10.1177/0956797617741719

Strevens, M. (2017). *Notes on Bayesian confirmation theory*. Retrieved from https://www.strevens.org/bct/BCT.pdf

Swafford, M., & Anderson, R. (2020). Addressing the gender gap: Women's perceived barriers to pursuing STEM careers. *Journal of Research in Technical Careers*, *4*(1).

https://doi.org/10.9741/2578-2118.1070

World Economic Forum. (2023). *Global Gender Gap Report 2023*. https://www3.weforum.org/

Yazilitas, D., Saharso, S., de Vries, G. C., & Svensson, J. S. (2017). The postmodern perfectionist, the pragmatic hedonist and the materialist maximalist: Understanding high school students' profile choices towards or away from mathematics, science and technology (MST) fields in the Netherlands. *Gender and Education*, *29*(7), 831-849. https://doi.org/10.1080/09540253.2016.1166185

Appendix

Bayesian Results Table

Row	Variables	Statistics	Evidence strength
1	Gender gap in intraindividual science strength; GGGI	BF_{10} = 15.22, M = .661	**Strong evidence in favor of effect**
1.1	Gender gap in intraindividual science strength; GGGI (linear, quadratic, and cubic models)	Cubic (best model) BF_M = 9.011, R^2 = .214 Linear: β = -.008, SD = .020, CI 95% = [-.061, .031], $BF_{inclusion}$ = .656 Quadratic: β = -.003, SD = .006, CI 95% = [-.018, .007], $BF_{inclusion}$ = .583 Cubic: β = -.016, SD = .006, CI 95% = [0.00, .028], $BF_{inclusion}$ = 41.457 All VIFs < 5	**Moderate evidence**
2	Gender gap in intraindividual math strength; GGGI	BF_{10} = .297	Inconclusive
3	Gender gap in intraindividual reading strength; GGGI	BF_{10} = .463	Inconclusive
4	Gender gap in intraindividual science strength; gender gap in university enrollment	BF_{10} = .392	Inconclusive
5	Gender gap in intraindividual science strength; gender gap in mean years of schooling	BF_{10} = .547	Inconclusive

6	Gender gap in intraindividual science strength; gender gap in expected years of schooling	BF_{10} = .391	Inconclusive
7	Gender gap in STEM aspirations; GGGI	BF_{10} = .344	Inconclusive
8[a]	Gender gap in STEM aspirations; gender gap in expected years of schooling	BF_{10} = .388	**Inconclusive**
9	Gender gap in STEM aspirations; gender gap in university enrollment	BF_{10} = .344	Inconclusive
10	Gender gap in STEM aspirations; gender gap in mean years of schooling	BF_{10} = .431	Inconclusive
11	Gender gap in intraindividual science strength; GGGI (linear, quadratic, and cubic models)	All three terms (best model): BF_M = 3.195, R^2 = .333, P(M|data) = .450 Quadratic + Cubic: BF_M = 4.241, R^2 = .312, P(M|data) = .278 Linear: β = .029, *SD* = .030, 95% CI [-.007, .089], $BF_{inclusion}$ = 2.495 Quadratic: β = -.028, *SD* = .013, 95% CI [-.049, .000], $BF_{inclusion}$ = 14.396 Cubic: β = .011, *SD* = .008, 95% CI [-.0003, .025], $BF_{inclusion}$ = 4.360 Null model: BF_M = .002 All VIFs < 5	**Moderate evidence**
12	Gender gap in intraindividual science strength; gender gap in university enrollment	BF_{10} = .307	Inconclusive
13[b]	Gender gap in intraindividual science strength; gender gap in mean years of schooling	BF_{10} = 3.859	**Moderate evidence in favor of effect**

14	Gender gap in intraindividual science strength; gender gap in expected years of schooling	BF_{10} = .307	Inconclusive
15	Gender gap in STEM aspirations; GGGI	BF_{10} = .333	Inconclusive
16[c]	Gender gap in STEM aspirations; gender gap in expected years of schooling	BF_{10} = .347	**Inconclusive**
17	Gender gap in STEM aspirations; gender gap in university enrollment	BF_{10} = .332	Inconclusive
18	Gender gap in STEM aspirations; gender gap in mean years of schooling	BF_{10} = .502	Inconclusive
19	GGGI; STEM graduation propensity	BF_{10} = 13.931, *M* = -.38, *SD* = .17, 95% CI [-.63, .00], R^2 = .176	**Strong evidence in favor of effect**
20	GGGI; Actual % of women in STEM degrees	BF_{10} = .350, *M* = -.016, *SD* = .13, 95% CI [-.36, .31], R^2 = .003	Inconclusive
21	Mean years of schooling; STEM graduation propensity	BF_{10} = 1.400, *M* = -.13, *SD* = .15, 95% CI [-.45, .032], R^2 = .174	**Inconclusive**
22	Mean years of schooling; Actual % of women in STEM degrees	BF_{10} = 14.502, *M* = .63, *SD* = .24, 95% CI [.00, .98], R^2 = .619	**Strong evidence in favor of effect**
23	Expected years of schooling; STEM graduation propensity	BF_{10} = .345, *M* = -.001, *SD* = .004, 95% CI [-.002, .014], R^2 = .010	Inconclusive
24	Expected years of schooling; Actual % of women in STEM	BF_{10} = 3.615, *M* = .09, *SD* = .06, 95% CI [.00, .19], R^2 = .169	**Moderate evidence in favor of effect**

25	University enrolment; STEM graduation propensity	BF_{10} = .324, *M* = .000, *SD* = .003, 95% CI [-.004, .010], R^2 = .007	Inconclusive
26	University enrolment; Actual % of women in STEM	BF_{10} = 4.111, *M* = .09, *SD* = .06, 95% CI [.00, .20], R^2 = .177	**Moderate evidence in favor of effect**
27	STEM graduation propensity; Actual % of women in STEM; Gender gaps in aspirations; Gender gaps in science strength	Propensity + gender gaps in science strength: BF_{10} = .369 Propensity + gender gaps in aspirations: BF_{10} = .512 Actual % + gender gaps in science strength: BF_{10} = .348 Actual % + gender gaps in aspirations: BF_{10} = .439	Inconclusive

Note. Row 1–6 represent findings on gender differences in relative academic strength (PISA2018); Row 7–10 represent findings on gender differences in STEM aspirations (PISA2018); Row 11–14 represent findings on gender differences in relative academic strength (PISA2015); Row 15–18 represent findings on gender differences in STEM aspirations (PISA2015); Row 19–27 represent findings on gender differences in STEM graduation (PISA2018). GGGI = Global Gender Gap Index. The highlighted rows in the 'Evidence strength' column signify a difference between G24's results and the results found in this study (e.g., row 8 was significant in G24. In the Bayesian reanalysis, the strength of the evidence is different from the G24 result); The bold rows resemble the significant rows in G24 (e.g., row 1 was significant in G24, and had a similar result in the Bayesian reanalysis).

[a] Countries not included in G24: ARG, ESP, ISL, ISR, JOR, JPN, MNE, RUS, TUR; (Countries included in G24, but not in the current study: DOM, PAN, PER, URY). [b] Countries not included in G24: DOM, HKG; (Countries included in G24, but not in the current study: ALB, ARE,

CAN, DZA, GBR, GEO, HRV, IRL, JOR, JPN, LBN, MKD, MNE, POL, QAT, RUS, THA, TTO, TUN, VNM). [c] Countries not included in G24: DOM, HKG, MAC; (Countries included in G24, but not in the current study: ARE, DZA, GRC, PER, SGP, THA, TTO, TUN, URY). BF_{10} = Bayes Factor comparing H_1 to H_0. BF_M = Bayes Factor comparing each model to all the other models. P(M|data) = posterior probability of each model given the observed data. $BF_{inclusion}$ = Evidence for including a predictor across all tested models.