Illusory gender-equality paradox, math self-concept, and frame-of-reference Effect : New integrative explanations for multiple paradoxes

Marsh, Herbert W., Parker, Philip D., Guo, Jiesi, Basarkod, Geetanjali, Niepel, Christoph and Van Zanden, Brooke

Illusory Gender-Equality Paradox, Math Self-concept, and Frame-of-Reference Effects: New Integrative Explanations for multiple Paradoxes

Herbert W. Marsh, Philip D. Parker, Jiesi Guo, Geetanjali Basarkod

Institute for Positive Psychology and Education, Australian Catholic University

Christoph Niepel

University of Luxembourg

Brooke Van Zanden

Institute for Positive Psychology and Education, Australian Catholic University

Corresponding Author: Herbert W. Marsh, Institute for Positive Psychology and Education (IPPE), Level 10, 33 Berry Street, North Sydney 2060, Herb.Marsh@acu.edu.au; +61 2 9701 4626

Philip D. Parker, IPPE, North Sydney Campus, Level 9, 33 Berry Street, North Sydney 2060. PO Box 968, North Sydney NSW 2059. E: Philip.Parker@acu.edu.au
Jiesi Guo, IPPE, North Sydney Campus, Level 9, 33 Berry Street, North Sydney 2060. PO Box 968, North Sydney NSW 2059. E: Jiesi.Guo@acu.edu.au
Geetanjali Basarkod, IPPE, North Sydney Campus, Level 9, 33 Berry Street, North Sydney 2060. PO Box 968, North Sydney NSW 2059. E: Geetanjali.Basarkod@acu.edu.au
Christoph Niepel, Department of Behavioural and Cognitive Sciences, University of Luxembourg, 11, Porte des Sciences; 4366 Esch-sur-Alzette, Luxembourg. E: christoph.niepel@uni.lu
Brooke Van Zanden, IPPE, North Sydney Campus, Level 9, 33 Berry Street, North Sydney 2060. PO Box 968, North Sydney NSW 2059.

This research was supported by a grant from the Luxembourg National Research Fund (FNR) to Christoph Niepel (C16/SC/11333571).

Abstract

Gender-Equality Paradoxes (GEPs) posit that gender gaps in math self-concepts (MSCs) are larger — not smaller—in countries with greater gender-equality. These paradoxical results suggest that efforts to improve gender-equality might be counter-productive. However, we show that this currently popular explanation of gender differences is an illusory, epi-phenomenon (485,490 students, 18,292 schools, 68 countries/regions). Between-country (absolute) measures of gender-equality are confounded with achievement and socioeconomic-status; tiny GEPs disappear when controlling achievement and socioeconomic-status. Critically, even without controls, GEPs are not supported when using true gender-gap measures—within-country (relative) female-male differences, that hold many confounds constant. This absolute/relative-gap distinction is more important than the composite/domain-specific distinction for understanding why even tiny GEPs are illusory. Recent developments in academic self-concept theory are relevant to GEPs and gender differences, but also explain other, related paradoxes. The big-fish little pond effect posits that attending schools with high school-average math achievements leads to lower MSCs. Extending this theoretical model to the country-level, we show that countries with high country-average math achievements also have lower MSCs. Dimensional comparison theory predicts that MSCs are positively predicted by math achievements but negatively predicted by verbal achievements. Extending this theoretical model, we show that girls' low MSCs are due more to girls' high verbal achievements that detract from their MSCs than to their low math achievements. In support of the pan-human wide generalizability of our findings, our cross-national results generalize over 68 country/regions as well as multiple math self-belief constructs (self-efficacy, anxiety, interest, utility, future plans) and multiple gender-equality measures.

Keywords: Math self-beliefs, Gender differences, Gender-Equality Paradox, cross-national comparisons, Social and dimensional comparison theories

Significance Summary

Paradoxically, recent research suggests that countries with greater gender equality have even bigger gender-gaps favoring boys' math self-beliefs. This finding is worrisome for efforts to increase gender-equality. However, we show that this Gender-Equality Paradox is illusory, an epi-phenomenon explained by controlling achievement and socioeconomic-status, and using more appropriate gender-equality measures. Dimensional comparison theory predicts that math self-beliefs are positively predicted by math achievement but negatively predicted by verbal achievement. Extending this theoretical model, we show that girls' low math self-beliefs are due more to girls' high verbal achievements that detract from MSC than to their low math achievements. In support of the wide generalizability of our findings, our cross-national results generalize over 68 country/regions and multiple constructs (self-concept, self-efficacy, anxiety, interest, utility, and future plans).

Gender-Equality Paradoxes (GEPs) is based on the finding that gender gaps in math self-concepts (MSCs) are larger— not smaller — in countries with greater gender-equality; paradoxically, girls are more disadvantaged in more gender-equal countries (Baker & Jones, 1993; Guo, et al., 2019; Marsh, Van Zanden, et al., 2019; Niepel, 2019; Stoet & Geary, 2018; Stoet, Bailey & Moore, 2016). MSC is an important construct that is reciprocally related to academic achievement, predicts coursework selection and long-term educational attainment, and contributes to gender imbalances in STEM disciplines (Marsh, 2007; Marsh, Van Zanden, et al., 2019). Similarly, Charles and Bradley (2009) found girls' affinity for math was higher in developing countries than advanced industrial countries. Support for GEPs (greater disadvantage for girls in more gender-equal countries) is not only paradoxical, but also calls into question international efforts to enhance gender equality. However, we posit that this GEP is illusory, an epi-phenomenon, and explore alternative explanations based on how measures of gender equality are constructed, adding appropriate controls for achievement and socioeconomic-status, and applying recent theoretical advances in academic self-concept theory.

Gender-Equality Paradox: Country Level Measures of Gender-Equality

The measurement of gender equality at the country-level is highly contentious with hundreds of measures largely based on ad hoc, atheoretical rationales (Else-Quest & Hamilton, 2018; Hawken & Munck, 2013; Klasen, 2018; Stoet & Gery, 2019). Else-Quest and Hamilton (2018) provide a historical perspective on the development of these measures, starting with the Fourth World Conference on Women, hosted in 1995 by the United Nation Commission on the Status of Women to promote improved statistical methodology and appropriate data for the economic, social, cultural and political development of women. In the period since this conference, many gender-equality measures have been used, including those published by the United Nations on their website (The World's Women; *http://unstats.un.org/unsd/gender/worldswomen.html*).

Common-sense and international efforts to reduce gender inequality are based on the premise that improving gender equality will lead to reductions in the sizes of gender gaps disadvantaging girls – the gender stratification hypothesis. Particularly in relation to gender-stratification and hypotheses based on GEP, Else-Quest et al (2010; Else-Quest & Hamilton, 2018; also see Niepel, 2019; Stoet &

Geary, 2016) highlighted the distinction between composite indices (weighted-averages of multiple indicators) and domain-specific single indicators of gender equality. Composite indices combine various differentially weighted indicators related broadly to human development that might differ substantially in their relevance to a particular study. These composite indictors are typically closely related to the Human Development Index indices (HDI) that focus on gender gaps in relation to women; some are even truncated at 1.0 (on a 0-to-1 scale) if women score higher than men (e.g., the Global Gender Gap Index, GGGI), so that they do not reflect gender differences in favor of women, whereas others measure equality in terms of absolute deviations that do not differentiate gender differences in favor of men or women.

Domain-specific measures represent a single indicator of national gender equality rather than a composite of different domain specific measures. Else-Quest and Grabe (2012) specifically recommended the use of domain-specific rather than composite measures to better identify or assess processes underling gender differences. Recognizing that gender inequality is a multidimensional construct, Else-Quest and Hamilton (2018) argued that it is best represented by domain-specific measures most relevant to particular studies, whereas composite measures are not internally consistent, difficult to interpret, disguise the complexity of gender equality, and have little meaning. In subsequent research, this composite/domain-specific distinction has been widely acknowledged (but also see related discussion by Bollen & Bauldry, 2011, on the difference between reflective measures and formative measures that are intended to provide an index of discrete measures rather than a summary of internally consistent indicators).

Based on traditional composite measures (Global Empowerment Measure; and Global Gender Gap Index, GGGI; see Table 1 and Appendix 1 for a list of definitions of key constructs used in the present investigation), Else-Quest et al. found support for GEPs in relation to math self-beliefs, but no support for the stratification hypothesis. However, based on a set of seven domain-specific indicators of gender equality, their results were mixed; most effects were non-significant but in some cases the results favored their gender stratification hypothesis. Stoet and Geary (2016), based on the composite GGGI, found that more gender-equal and economically-advanced countries had larger gender differences in math anxiety—consistent with the GEP. Niepel et al. (2019), based on a domain-specific measure (percentage of scientific jobs held by women), found support for the stratification hypothesis. This pattern of results apparently supports the importance of the composite/domain-specific distinction as argued by Else-Quest and Hamilton (2018), but alternative explanations exist.

Another critical difference is whether measures vary on an absolute metric that facilitates comparisons across countries (e.g., Human Development Index, HDI) or within-country relative scores that reflect the juxtaposition of gender differences within a given country (i.e., female/male ratios or female-male differences; see Table 1 and Appendix 1). Domain-specific absolute measures (e.g., country-level achievement) and composite scores such as the widely used GGGI, are based on a complex mixture of absolute and relative differences, non-linear transformations of component scores, and truncation of scores in favor of women that potentially confound gender equality with overall country levels of development. In contrast, within-country differences hold country level differences constant and focus on male-female differences within each country. This relative/absolute distinction is important because the domain-specific measures used by Else-Quest et al. (2014) and Niepel et al. (2019) are all relative measures, whereas the Global Empowerment Measure and GGGI composites confound the composite/domain-specific and relative/absolute distinctions. However, it is possible to construct domain-specific and composite measures that are either absolute or relative (see Table 1), providing a test of the relative/absolute vs. composite/domain-specific distinctions.

Although a comprehensive review of gender-equality measures is beyond the scope of the present investigation, we use a variety of measures to illustrate these distinctions (see Table1 and Appendix 1). In relation to the measures listed in Table 1, we note that all of them can be used to evaluate GEPs, even though many do not actually measure gender equality per se. Thus, for example, if gender differences in favor of men are significantly greater in more economically advanced countries based on various measures of economic development (e.g., HDI, OECD—Organisation for Economic Co-operation and Development countries, SES—socioeconomic status, gross domestic product) or measures related to economic development, then there is support for the GEP.

Based on the finding that most countries where girls score higher than boys in math and science are Middle Eastern Islamic countries (Marsh, Abduljabbar et al., 2014), we include the percent of Muslims as one of our country-level contextual variable. Following discussion by Else-

Quest et al. (2014) on cultural differences in gender stratification, we also include country-level measures based on Hofestede's (1984) cultural differences (see Appendix 1).

We also introduce new contextual measures based on the variables available in the Programme for International Student Assessment (PISA) database collected in 2012. These indicators (country-level achievement and SES, and relative ratio and difference measures based on these measures) are well documented and available in the public domain (noting that PISA is the most comprehensive and rigorous international data used to assess student performance). We relate these indicators to student, family, institutional, and country factors that can help to explain differences in performance (https://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm).

Of relevance, we juxtapose composite measures (e.g., HDI) and domain-specific measures (e.g., university enrollment of women, country-average achievement) that are represented as both relative and absolute measures (e.g., HDI, within-country female-male gender differences in HDI, within country female/male HDI ratios). In this way we unconfound the relative/absolute and composite/domain-specific distinctions which has not been made explicit in previous research. We also note that most so-called measures of gender gaps are really female/male ratios that potentially confound the effects of the ratio (a multiplicative, cross-product "interaction" term) with the first-order ("main") which complicates interpretation, an issue well-known to psychologists in relation to testing/interpreting interaction effects in regression analyses (e.g., Cohen, West & Aiken, 2014). For this reason, as well as because they are easier to interpret (see related discussion by Klasen, 2018), we have also included true gender gap measures that are simply within-country male-female differences.

Academic Self-concept: Gender Differences and Theoretical Models

Related to the issue of gender equality is the question of why, internationally, girls have lower MSCs than boys, even after controlling for gender differences in math achievement and country-level gender inequality—a dilemma that can be explained in part by application of academic self-concept (ASC) theory that has been largely ignored in GEP studies. Here we briefly review gender differences in MSC and two theoretical models (social comparison and dimensional comparison theories) that contribute to understanding these gender differences.

Gender Differences in Math Self-Concept

Gender differences in self-concept have long been a topic of interest (see Wylie, 1979), highlighted by findings (e.g., Marsh, 1989) that gender-stereotypic differences in multiple dimensions of self-concept (e.g., boys higher in MSC but lower in verbal self-concept) were consistent over the pre-adolescent to young-adult period. However, Hyde (2007) noted most gender-difference research is based on middle-class white college students from Western countries, and called for more cross-national studies of gender differences and how they vary with country-level gender equality. In their cross-national PISA study, Else-Quest, et al (2010) found only small differences in math achievement (ES < .15), but much larger gender differences in favor of males for MSC (ES = .33). In support of GEPs and contradicting their gender stratification hypothesis, Else-Quest et al. reported that gender gaps in favor of boys for MSC, self-efficacy, extrinsic motivation, intrinsic motivation, and anxiety tended to be larger—not smaller—in countries with higher levels of gender equality (also see Stoet et al., 2016; 2018). We contend that these findings are an epi-phenomenon due to problems in the measurement of gender-equality and development at the country level (see earlier discussion), a lack of control for relevant country-level differences that confound the results, and failure to incorporate recent advances in ASC theory that provide alternative explanations for the results. Here we outline ASC theory as it relates to our study (for further discussion see Marsh, 2007; 2016).

Social Comparison Theory

William James (1890/1983) recognized that self-concept is based on objective accomplishments evaluated in relation to frames-of-reference, a perspective also central to Festinger's (1954) social comparison theory. Focusing on ASC in educational contexts, Marsh (1984; see also Marsh & Parker, 1984; Marsh & Seaton, 2015) proposed the big-fish-little-pond effect (BFLPE), in which students have lower ASCs in the context of high-ability students, but higher ASCs in the context of low-ability students. In the theoretical model underpinning the BFLPE (Figure 1B), the effects of student-level achievement on ASC are positive, but the effects of school-average achievement are negative. Extending the BFLPE, country-level achievement also has a negative effect on ASC (Marsh, 2016; see Figure 1B), in addition to the negative effect of school-average achievement. This provides a theoretical explanation to the paradoxical cross-country effect (students in high achieving countries have lower ASCs after controlling for individual achievement) that may be related to GEPs.

Dimensional Comparison Theory (DCT)

DCT (Marsh, 1986, 2007; Möller & Marsh, 2013), posits that academic accomplishments in one domain serve as a basis of comparison for the formation of self-concepts in other domains. DCT predicts that MSC is positively predicted by math achievement, but negatively predicted by verbal achievement (Figure 1A). Hence, DCT predicts that high verbal ability detracts from MSC. The generalizability of these predictions was supported over 26 countries using PISA data (Marsh & Hau, 2004) and in subsequent meta-analyses (Möller, et al., 2009; Huang, 2011).

Building on Marsh's frame of reference research (Marsh, 1986; 2007; also see Eccles, 2009) and their expectancy value theory, Wang, Eccles and Kenny (2013, pp. 770-771; also see Wang & Degol, 2017) proposed that "individuals with high math and moderate verbal ability have higher math-ability self-concepts than individuals with high math and high verbal ability" (also see related discussion by Ceci, et al. 2014; Stoet, et al. 2013; 2018). For a nationally representative sample of high school students, they found that by age 33, "mathematically capable individuals who also had high verbal skills were less likely to pursue STEM careers than were individuals who had high math skills but moderate verbal skills" (p. 770). Although Stoet, et al. (2018) used a similar logic to Wang et al., they did not formally test predictions based on DCT. Additionally, their use of difference scores confounded positive effects of math achievement with the negative effects of verbal achievement. More recently, Marsh, Van Zanden, et al. (2019) reported that girls' higher verbal achievement detracted from their math self-concept, self-efficacy, interest, and value (and increased anxiety) independent of their mathematics achievement. Integrating subsequent advances in DCT with these results, we posit that gender differences in MSC are substantially due to girls' higher verbal achievement (that detracts from MSC) as well as girls' lower math achievement – a seemingly paradoxical prediction that follows from DCT and is consistent with previous research.

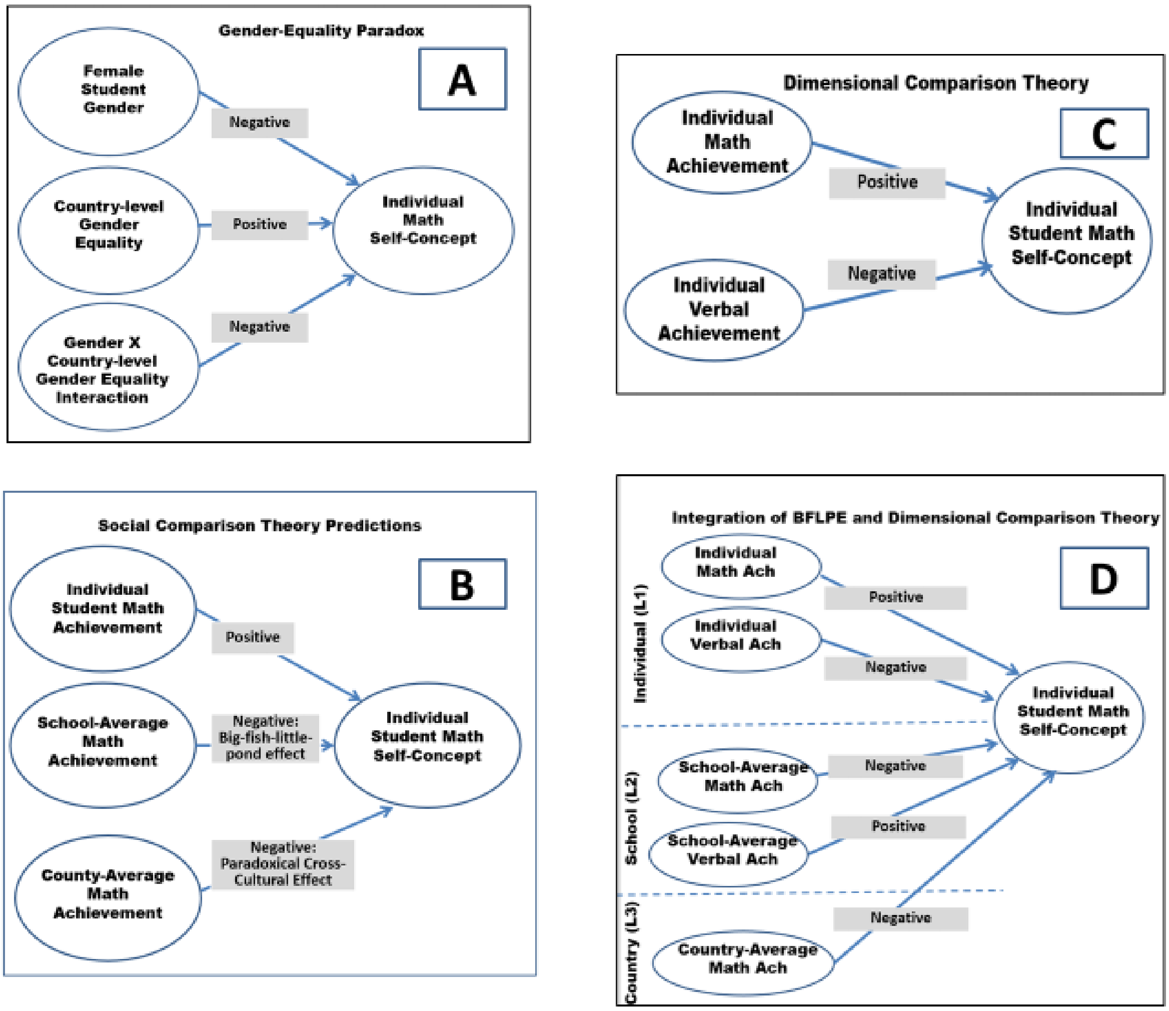
Gender-Equality Paradox
A
Female Student Gender
Negative
Country-level Gender Equality
Positive
Individual Math Self-Concept
Gender X Country-level Gender Equality Interaction
Negative
Dimensional Comparison Theory
C
Individual Math Achievement
Positive
Individual Student Math Self-Concept
Individual Verbal Achievement
Negative
Social Comparison Theory Predictions
B
Individual Student Math Achievement
Positive
School-Average Math Achievement
Negative: Big-fish-little-pond effect
Individual Student Math Self-Concept
County-Average Math Achievement
Negative: Paradoxical Cross-Cultural Effect
Integration of BFLPE and Dimensional Comparison Theory
D
Individual (L1)
School (L2)
Country (L3)
Individual Math Ach
Positive
Individual Verbal Ach
Negative
Individual Student Math Self-Concept
School-Average Math Ach
Negative
School-Average Verbal Ach
Positive
Country-Average Math Ach
Negative

Figure 1.

(A) Gender-equality Paradox (GEP). The GEP posits that the negative effect of gender (female) on MSC is more negative in countries with higher levels of gender equality. This opposes the predictions of the gender stratification hypothesis that posits that the interaction should be positive (i.e., gender differences in favor of males will be smaller in countries with greater gender equality).

(B) Social Comparison Theory Predictions. Big-fish-little-pond effect (BFLPE): Predicts that the effect of school-average achievement on math self-concept is negative. (C) Paradoxical cross-cultural effect: predicts that the effect of country-average math achievement on math self-concept is negative.

(C) Dimensional Comparison Theory (DCT): Predicts that the effect on math self-concept is positive for math achievement but negative for verbal achievement (and that the effect on verbal self-concept is positive for verbal achievement but negative for math achievement).

(D) Integration of Social and Dimensional Comparison Theory. The Big-fish-little-pond compensatory effect (BFLPE-CE) has the same predictions as social comparison theory (Figure 1B) and dimensional comparison theory (Figure 1C). The new predictions are for the effects for school-average verbal achievements. Integrating the logic of the DCT and BFLPE models regarding school-average achievement, the effects are predicted to be in the opposite direction of those for individual achievement; negative for school-average math achievement (BFLPE), but positive for school-average (BFLPE-CE)

Integration of Social Comparison and Dimensional Comparison Theories

BFLPE studies typically focus on one domain (e.g., math) at multiple levels (e.g., L1, student; L2, school), whilst DCT studies typically focus on multiple domains (e.g., math and verbal) at a single level (L1, student). Juxtaposing the two theoretical models, Marsh (1986; 1994; Marsh, Parker & Craven, 2015; Parker, Marsh, et al., 2013) posited BFLPE compensatory effects (BFLP-CE), such that school-average math achievement has a negative effect on MSC (the BFLPE), while school-average verbal ability should have a small, compensatory positive effect on MSC (BFLP-CE, Figure 1D; see Parker et al., 2013; Pinxten et al., 2015).

The Present Investigation: Statement of Research Hypotheses

For present purposes we posit a set of research hypotheses and questions that guide our analyses and presentation of results. We test these results based on the publicly available PISA2012 data, which includes nationally representative samples of 15-year-old students from 68 countries/regions.

1. **Gender Equality Paradox (GEP).** So-called GEPs are small in size and largely eliminated by controlling for measures of country-level achievement and SES that are confounded with absolute measures of gender-equality and development. There is no support for GEPs using relative measures of gender-equality (female/male ratios or differences within each country) that hold country-level indicators constant.
2. **Gender Differences Paradox.** Consistent with dimensional comparison theory (DCT) predictions, gender differences in MSC are substantially reduced by controlling girls' higher levels of verbal achievement, but reduced little or not at all by controlling girls' lower levels of math achievement.
3. **Big-Fish-Little-Pond Effect (BFLPE).** After controlling for individual (L1) math achievement and other variables in the model, the first-order effect of L2 school-average math achievement on MSCs is negative (the BFLPE based on social comparison frame-of-reference effects). For this and each of the subsequent hypotheses, we leave as a research question whether this effect interacts with gender.
4. **Paradoxical Country-Level Frame-of-Reference Effects.** Consistent with extensions of the BFLPE, after controlling for L1 and L2 math achievement and in the context of the GEP

analyses, the first-order effect of L3 (country-level) math achievement on MSCs is negative (paradoxical cross-national effect). Consistent with Hypothesis 1, there are small but statistically significant interaction effects (in support of apparent GEPs) for all the absolute and mixed contextual variables—whether composite or domain specific. In each case, the direction of these interaction effect is such that gender differences are larger in more advantaged countries (noting that for Gender Inequality Index, % Muslim, and Power Distributions, higher scores are associated with greater levels of disadvantage in relation to, for example, to HDI—see correlations with HDI in Table 1). It is, however, important to emphasize that even these effects apparently in support of GEPs are tiny in size; only the effect of OECD (ES = .12) is greater than the .1 value that Else-Quest et al., 2014 claims is negligible and none even approach the value of .2 that is considered to be a small ES. Absolute gender-equality measures sometimes seem to support GEPs as these absolute gender-equality measures are substantially correlated with country-level indicators such as achievement and SES.

5. **Dimensional Comparison Effects (DCT).** After controlling for math achievement and in the context of the GEP analyses, the first-order effect of L1 verbal achievement on MSC is negative (dimensional comparison effect—also see hypothesis 2).
6. **Social and Dimensional Comparison Integration.** After controlling for math achievement and L1 verbal achievement, and in the context of the GEP analyses, the first-order effect of L2 verbal achievement on MSC is positive effect (compensatory BFLPE, based on the integration of social and DCT). This follows from the prediction that the effect of L2 math achievement is negative (BFLPE, hypothesis 3) and DCT prediction that the effects of verbal achievement on math self-concept are opposite those of math achievement.
7. **Generalizability.** Support for all hypotheses based on MSC and the GGGI gender-equality indicator generalize in separate analyses of other math psychological variables (math self-efficacy, anxiety, extrinsic and intrinsic motivation, and future plans) and over different country-level indicators of gender equality and economic development. Nevertheless, given the frame-of-reference basis of the predictions based on ASC research and theory, the sizes of effects are expected to be larger for MSC than for the other variables.

Method and Materials

Data Accessibility

Data used in the present investigation are publicly available through the OECD-PISA website (https://www.oecd.org/pisa/pisaproducts/). Available through this website is a link for downloading the PISA2012 data used in this study, technical reports providing a detailed description of the implementation of the survey and psychometric evaluation of the survey responses, an extensive set of theoretical, substantive, and policy oriented publications based on the PISA survey. Also included is code (e.g., an SPSS syntax file) for reading and analyzing the data. The specific variables used as part of the present investigation are listed in the Appendix of this article, and specific details on the analyses of these data are presented below. Although there are a huge number of publications based on PISA data, there is an extensive archive of working papers available through the PISA website and the most relevant publications are reviewed in this article (as well as showing how our study advances on this previous research). The Australian Catholic University exempts from the need for ethics approval publicly available data in the public domain—including secondary data analyses based on PISA data.

PISA Sample and Variables

For the PISA2012 data used here (485,490 fifteen-year-old students, 18,292 schools, 68 countries/regions), the primary focus is on math. See Supplemental Table S1 for a listing of the 68 countries and the number of students and schools from each country.

Students completed paper-and-pencil tests to assess their knowledge and skills in reading, math, and science. In addition, each student completed a questionnaire assessing student and family background variables, and a variety of psychosocial variables, with a focus on math. These data are publicly available through the OECD-PISA website (https://www.oecd.org/pisa/pisaproducts/), as well as in the extensive documentation, and technical reports on collected variables. Data were collected using a complex two-stage sampling design, and were nationally representative samples, after using the appropriate survey weights. The central variables considered here are as follows:

MSC, the main dependent variable, was measured with five items (e.g., “I learn mathematics quickly”). Based on item response theory, it was represented as the Weighted Likelihood Estimate provided with the PISA data, as recommended in the PISA manual (OECD, 2014). In addition to MSC, other math-specific constructs (e.g., self-efficacy, anxiety, interest, instrument motivation, and future plans) are considered (see OECD, 2014, for a more detailed description including the wording of the items and psychometric support for the measurement of these constructs). Gender-equality indices, country-level contextual variables, and other math-specific constructs (e.g., self-efficacy, anxiety, interest, instrument motivation, and future plans) are summarized in Appendix 1 (also see OECD, 2014, for more detail).

In the PISA database, math, reading, and science achievement are based on achievement tests completed by each student. School- and country-average achievement were based on individual student (L1) measures of achievement aggregated to the school-average (L2) and country-average (L3) levels, whilst the interaction between L1 and L2 math achievement (Mach and L2Mach) was represented as the cross-product of those two variables (L1xL2 math achievement).

Analysis

PISA2012 sampling design and implications for multiple imputation analysis. In the PISA2012 database provided to users, math, reading, and science achievements at the individual student level are each represented by a set of five plausible values designed to prevent biased population estimates, control for measurement error, and facilitate secondary data analysis such as the present investigation (see OECD, 2014). As outlined in the PISA Technical Report that was the basis of analyses in the present investigation (see below), appropriate analysis with plausible values is to undertake each analysis five times, once with each plausible value and to then combine the results using Rubin's (1987) approach (for further detail on construction of these plausible values see OECD, 2014). There were no missing data for the plausible values of math achievement.

As noted in the PISA2012 Technical Report (OECD, 2014), whereas rotation of cognitive test items has been used regularly in PISA data collections, 2012 is the first time this strategy has been used for student context surveys. This was done to increase the content coverage, whilst maintaining the amount of time needed to complete the survey. There were three survey forms, each of which contained a common set of items and a rotated section. In the rotated section, students completed 2/3 of the rotated items, such that allocation was based on the use of intact scales that were balanced in terms of correlations with performance. Because responses based on this strategy were purely missing completely at random, they were appropriately handled using multiple imputation. In particular, math self-concept was rotated, so that approximately 1/3 of these variables were missing by design. For the present purposes, these missing values were appropriately handled by the multiple imputation strategy.

Because of the design of the PISA2012—the provision of five plausible values to represent achievement and missing by design for survey items—we used multiple imputation (Rubin, 1987) to deal with the missing data. Using a large imputation model, five imputed datasets were created using Markov chain Monte Carlo (MCMC) imputation, including dummy variables to represent the 68 countries. The decision to use multiple imputation was based on the need to include many auxiliary variables in the imputation model, the need to use plausible values for the achievement test that require an analytical strategy akin to the analysis of multiply imputed datasets, and also the rotation strategy first introduced in PISA2012 for the student survey, with non-cognitive items. To account for the five plausible values of achievement provided as part of the PISA database, each of the five imputations included one of these five plausible values, and results from the five imputations were combined using Rubin's (1987) approach.

Multilevel modeling (MlwiN; Rasbash, Steele, Browne, & Prosser, 2004; also see Marsh, 2016) was used to accommodate the three-level hierarchical PISA structure: students (L1) nested within schools (L2), and schools nested within countries (L3). Cases were weighted by the weighting variable provided as part of the PISA database (OECD, 2014). The fixed effects included individual achievement, school-average achievement, country-average achievement. Random effects included the intercepts at the three levels, but also country-level residual variances used to evaluate country-to-country variation in the critical frame-of-reference effects. Due in part to the PISA2012 design in which achievement is represented as five plausible values, we conducted analyses on five imputed data sets and combined them using Rubin's (1987) rules. In addition, we included two control variables (year in school and immigration status) that are not presented in the main text to conserve space (but are included in Tables in Supplemental Materials). Computer code used in the analysis of the data (macros for the MLwiN statistical package) are included in Supplemental Material 6.

Standardized metric. We standardized scores (*M* = 0, *SD* = 1) for all student, school, and country-level variables across the entire sample, to facilitate interpretations in relation to a standardized effect-size metric. However, dichotomous variables were scored 0 or 1 to facilitate interpretation in relation to a traditional effect-size metric. Values of 1 were assigned to girls (boys = 0) and students from OECD countries (non-OECD countries = 0). Variance components based on key constructs are presented in Supplemental Table 2. Although there is no clear agreement on what constitutes meaningful ESs, Else-Quest et al. (2014; Hyde, 2005) argue that gender-difference ESs < .1 are negligible and close to zero, even if statistically significant, and that ESs of .2, .5 and .8 are typically considered small, medium, and large, respectively. However, arguing for a gender similarity hypothesis, Hyde (2005) also notes that most gender differences are small or close to zero, but that gender differences in MSC (ES ≈ .3) tend to be larger.

Results

Tests of the Gender-Equality Paradox for Different Country-Level Indicators (Hypothesis 1)

In a series of multilevel models, we related country-level (L3) contextual (country-level gender-equality and development) variables, female gender, and their interaction to MSC. In these preliminary analyses, separate three-level analyses were done for each contextual variable, predicting individual MSC with individual gender (female), the country-level contextual variable, and its interaction with gender. The effect of gender on MSC (-.30; Table 1 with no contextual effects) is like that found in other studies (e.g., Else-Quest, et al., 2014). In these analyses, the critical effect is the interaction that represents the GEP—the extent to which gender differences vary with country-level contextual variables. GEPs are supported when the interaction is significant and favor girls (or favor boys to a lesser extent) in countries with lower levels of gender equality and development; interactions favoring girls in countries with higher levels of gender equality and development support a gender-stratification hypothesis.

Contextual variables can be roughly classified into between-country (absolute) measures and within-country (relative) measures that juxtapose results for males and females within each country, but can also be classified as composite and domain-specific measures. Nevertheless, there is a clear pattern of results (Table 1). As noted earlier (consistent with Hypothesis 1), there are small but statistically significant interaction effects (in support of apparent GEPs) for all the absolute and mixed contextual variables—whether composite or domain specific. In each case, the direction of these interaction effects is such that gender differences are larger in more advantaged countries (noting that for Gender Inequality Index, % Muslim, and Power Distributions higher scores are associated with greater levels of disadvantage in relation to, for example, HDI — see correlations with HDI in Table 1). It is, however, important to emphasize that even these effects apparently in support of GEPs are tiny in size; only the effect of OECD country (ES = .12) is greater than the .1 value that Else-Quest et al. (2014) claims is negligible, and none even approach the value of .2 that is considered to be a small ES.

In marked contrast to the absolute contextual variables, results using relative measures do not support GEPs; there are no significantly negative interaction effects for any of the (composite or domain-specific) relative measures. For example, for absolute HDI composite measures (total and separate for men and women) there is support for GEPs, but for relative HDI composite measures based on female-male differences or ratios, tests of GEPs are non-significant. Relatedly, for domain-specific measures of achievement there is support for GEPs based on total achievement (absolute measure averaged across boys and girls), but support for the gender-stratification hypothesis based on within-country gender differences (Total Achievement F-M Deviation in Table 1).

In support of our classification of measures, we note that absolute and mixed contextual variables (Table 1) are often highly correlated with HDI (|r| .37 - .97), whereas corresponding correlations with relative measures are much smaller (|r| .00 - .27) and mostly non-significant.

Table 1

Tests of the Gender-Equality Paradox: Prediction of Math Self-concept by Gender (female), Country-Level Contextual Variables, and the Gender-by-Context Interaction

Country Level Contextual Variables	Standardized Effects of												Nature of Context Variable			Correlation with HDI
	Female				Context Variable				Interaction							
	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI				
No Contextual Variables (68)	-.30	.02	-.33	-.27												
OECD Country (68)	-.24	.03	-.29	-.19	.00	.06	-.12	.12	-.12	.04	-.19	-.05	Paradox	Absolute	Dom-Specific	.70
Human Development Index (HDI, 60)	-.29	.02	-.32	-.26	.04	.03	-.02	.10	-.09	.02	-.12	-.06	Paradox	Absolute	Composite	1.00
HDI-females (56)	-.29	.02	-.32	-.26	.04	.03	-.02	.10	-.09	.02	-.12	-.06	Paradox	Absolute	Composite	.97
HDI-males (56)	-.29	.02	-.32	-.26	.03	.03	-.03	.09	-.08	.02	-.11	-.05	Paradox	Absolute	Composite	.95
Gender Inequality Index (53)	-.29	.02	-.33	-.25	.01	.03	-.06	.08	.08	.02	.04	.12	Paradox	Mixed	Composite	-.85
Global Equality Measure (59)	-.29	.02	-.32	-.26	.04	.03	-.02	.10	-.08	.02	-.11	-.05	Paradox	Mixed	Composite	.76
Reading Achievement (68)	-.29	.02	-.32	-.26	-.03	.03	-.09	.03	-.08	.02	-.11	-.05	Paradox	Absolute	Dom-Specific	.74
Total Achievement (68)	-.29	.02	-.32	-.26	-.03	.03	-.09	.03	-.08	.02	-.11	-.05	Paradox	Absolute	Dom-Specific	.74
HDI-Income loss (56)	-.29	.02	-.33	-.25	.03	.03	-.03	.09	-.08	.02	-.12	-.04	Paradox	Absolute	Composite	.91
Gross Domestic Product (63)	-.29	.02	-.33	-.25	.06	.03	.01	.11	-.07	.02	-.10	-.04	Paradox	Absolute	Composite	.74
Math Achievement (68)	-.29	.02	-.32	-.26	-.03	.03	-.09	.03	-.07	.02	-.10	-.04	Paradox	Absolute	Dom-Specific	.68
Science Achievement (68)	-.29	.02	-.32	-.26	-.03	.03	-.09	.03	-.07	.02	-.11	-.03	Paradox	Absolute	Dom-Specific	.65
Per-cent Muslim (64)	-.31	.02	-.34	-.28	.04	.03	-.01	.09	.07	.02	.04	.10	Paradox	Absolute	Dom-Specific	-.37
Global Gender Gap Index (55)	-.28	.02	-.32	-.24	.03	.03	-.02	.08	-.06	.02	-.09	-.03	Paradox	Mixed	Composite	.52
Power Distributions (55)	-.31	.02	-.35	-.27	-.07	.03	-.12	-.02	.06	.02	.03	.09	Paradox	Absolute	Dom-Specific	-.62
Individualism/collectivism (55)	-.31	.02	-.35	-.27	.06	.03	.00	.12	-.05	.02	-.09	-.01	Paradox	Absolute	Dom-Specific	.74
Socioeconomic status (SES, 68)	-.29	.02	-.33	-.25	.07	.03	.02	.13	-.05	.02	-.09	-.01	Paradox	Absolute	Composite	.86
HDI-F/M Ratio (60)	-.29	.02	-.33	-.25	-.02	.03	-.07	.03	-.01	.02	-.05	.03	Neither	Relative	Composite	.10
Gender Development Index (60)	-.29	.02	-.33	-.25	-.01	.03	-.06	.04	-.01	.02	-.05	.03	Neither	Relative	Composite	.17
% Female within Management (42)	-.30	.03	-.35	-.25	-.02	.03	-.07	.03	.00	.02	-.05	.05	Neither	Relative	Dom-Specific	-.17
HDI M-F Deviation (60)	-.29	.02	-.33	-.25	.01	.03	-.04	.07	.01	.02	-.03	.05	Neither	Relative	Composite	.27
Tertiary Enrollment F-M Deviation	-.29	.02	-.33	-.25	.04	.03	-.02	.10	.01	.02	-.03	.05	Neither	Relative	Dom-Specific	.25
Tertiary Enrollment F/M ratio	-.29	.02	-.33	-.25	.06	.03	-.01	.13	.03	.03	-.02	.08	Neither	Relative	Dom-Specific	.03
Per-cent Female within STEM Univ (50)	-.28	.02	-.32	-.24	-.01	.03	-.06	.04	.05	.02	.01	.09	Stratification	Relative	Dom-Specific	-.17
Total Achievement F-M Deviation (68)	-.30	.02	-.34	-.26	.01	.03	-.05	.07	.06	.02	.02	.10	Stratification	Relative	Dom-Specific	-.04

Note. Separate three-level analyses were done for each contextual variable, predicting individual math self-concept with individual gender (female), the country-level contextual variables listed above, and their interaction. The paradox is supported when the interaction is statistically significant and in a "paradoxical" direction. Contextual variables (numbers in parentheses are the number of countries for which the contextual variable was available) are classified as "relative" when gender differences are based on within country differences between men and women, "absolute" when measures are not based on gender differences or gender specific measures in relation to all countries, or "mixed" for composite measures that have relative and absolute components as well as complex transformation and truncation the complicate this distinction. We have also classified measures as domain-specific (Dom-Specific) and composite depending on whether the measure focuses on a specific variable or is an index that combines multiple domains. This interpretation is related to the extent to which each of the contextual measures is correlate to HDI, a widely used absolute measure country-level development. Control variables included gender, SES, individual student achievement in different domains, and a wide range of contextual variables. However, in relation to the purposes of this study these were important variable of interest so we have not specifically referred to them as control variables per se. Values in bold are statistically significant (p < .05). Est = estimate; SE =standard error; LCI = lower confidence interval; UCI = upper confidence interval;

Frame of Reference Effects: The Effects of SES and Achievement on MSC

In a series of models (Table 2), we systematically evaluate support for the set of hypotheses (also see Figure 1) with MSC as the outcome and GGGI as the contextual variable used to test the GEP. We note that GGGI is a composite measure for which there is support for the GEP here (Table 1) as well as other research (e.g., Else-Quest et al, 2010 analysis based on MSC from PISA2003). In these models, we include gender interactions that are a major focus of our study, and interpret the first order effects of hypothesized effects in relation to the full model including both covariates (GGGI and SES) and interactions. However, in supplemental materials (Supplemental Table 4), we also present results with no gender interactions (i.e. pure main-effect model) and models excluding both gender interactions and covariates. Because the interpretation of these supplemental analyses is essentially the same as for those based on Table 3, we do not focus on these in supplemental analyses.

Hypothesis 1. Model 1A shows an apparent GEP (the significantly negative GGGI-female interaction) that is small (ES = -.06). However, consistently with Hypothesis 1, this effect becomes non-significant when controlling L3-SES (Model 1D) and remains non-significant for all subsequent models (Models 1E-1K). This finding supports the contention that the apparent GEP in relation to GGGI is an artefact of GGGI being confounded with SES and achievement.

Hypothesis 2. Gender effects show a pattern of results in support of Hypothesis 2. Girls' lower MSCs (-.28, Model 1A) is unaffected by the introduction of SES (Models 1B-1D) and only marginally reduced by the introduction of math achievement at levels 1, 2 and 3 (Models 1E-1G; -.23 or -.22). However, consistently with Hypothesis 2 (based on DCT), the effect of gender is substantially reduced by the introduction of L1-verbal achievement (Model 1H; ES = -.06) and remains at this low level in subsequent models. Hence, the gender difference in MSC is more a function of girls' high levels of verbal achievement than their low levels of math achievement.

Table 2

Tests of Hypotheses 1-8: Multilevel Effects of Math Self-concept of Gender, Gender Equality (GGGI), SES, Math achievement, and Reading Achievement

	Model 1A				Model 1B				Model 1C				Model 1D				Model 1E				Model 1F				Hypo-thesis
Fixed Effects	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	
female	-.28	.02	-.32	-.24	-.28	.02	-.32	-.24	-.28	.02	-.32	-.24	-.28	.02	-.32	-.24	-.23	.02	-.26	-.20	-.22	.02	-.25	-.19	H2
L3-GGGI	*.03*	.06	-.09	.15	*.00*	.03	-.05	.05	*-.01*	.03	-.06	.04	*.01*	.03	-.05	.07	*-.02*	.06	-.13	.09	*.00*	.04	-.08	.08	
L3-GGGI.female	-.06	.02	-.11	-.01	-.05	.02	-.09	-.01	-.04	.02	-.08	.00	*-.04*	.02	-.08	.00	*-.03*	.02	-.07	.01	*-.04*	.02	-.07	-.01	H1
L1-SES					.13	.00	.12	.14	.13	.00	.12	.14	.13	.00	.12	.14	.06	.00	.05	.07	.05	.00	.04	.06	
SES.female					-.02	.00	-.03	-.01	*.00*	.00	-.01	.01	*.00*	.00	-.01	.01	-.02	.00	-.03	-.01	-.02	.00	-.03	-.01	
L2-SES									.02	.00	.01	.03	.02	.00	.01	.03	-.18	.00	-.19	-.17	-.03	.01	-.04	-.02	
L2-SES.female									-.05	.01	-.06	-.04	-.05	.01	-.06	-.04	-.04	.01	-.05	-.03	-.02	.01	-.03	-.01	
L3-SES													*-.03*	.03	-.09	.03	*.04*	.06	-.07	.15	*.01*	.04	-.07	.09	
L3-SES.female													*-.01*	.02	-.05	.03	*-.03*	.02	-.07	.01	*-.03*	.02	-.07	.01	
L1-MAch																	.44	.00	.43	.45	.49	.00	.48	.50	
L1-MAch.female																	.05	.00	.04	.06	.06	.00	.05	.07	
L2Mach																					-.24	.01	-.25	-.23	H3
L2Mach.female																					-.03	.01	-.04	-.02	
L3Mach																									H4
L3Mach.female																									
L1-RAch																									H5
L1-RAch.female																									
L2-RAch																									H6
L2-RAch.female																									
Random Effects																									
L3 Country Residual	.03				.03				.03				.03				.10				.05				
Female	.01				.01				.01				.01				.01				.01				
L2 School Residual	.03				.02				.02				.02				.04				.02				
Female	.01				.01				.01				.01				.01				.01				
L1 Student Residual	.90				.89				.89				.89				.77				.77				

Table 2 Continued

	Model 1G				Model 1H				Model 1I				Model 1J				Model 1K				Hypot-hesis
Fixed Effects	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	
female	-.22	.01	-.25	-.19	-.06	.01	-.09	-.03	-.07	.01	-.10	-.04	-.06	.01	-.09	-.03	-.06	.01	-.09	-.03	H2
L3-GGGI	.02	.03	-.04	.08	.02	.03	-.04	.08	.02	.03	-.04	.08	.02	.03	-.04	.08	.02	.03	-.04	.08	
L3-GGGI.female	-.03	.02	-.06	.00	-.03	.02	-.06	.00	-.03	.02	-.06	.00	-.02	.02	-.05	.01	-.02	.02	-.05	.01	H1
L1-SES	.05	.00	.04	.06	.06	.00	.05	.07	.06	.00	.05	.07	.06	.00	.05	.07	.06	.00	.05	.07	
SES.female	-.02	.00	-.03	-.01	-.02	.00	-.03	-.01	-.02	.00	-.03	-.01	-.02	.00	-.03	-.01	-.02	.00	-.03	-.01	
L2-SES	-.03	.01	-.04	-.02	-.01	.01	-.02	.00	-.02	.01	-.03	-.01	-.01	.01	-.02	.00	-.01	.01	-.02	.00	
L2-SES.female	-.02	.01	-.03	-.01	-.02	.01	-.03	-.01	-.02	.01	-.03	-.01	-.03	.01	-.04	-.02	-.03	.01	-.04	-.02	
L3-SES	.09	.03	.02	.16	.07	.03	.00	.14	.07	.03	.00	.14	.08	.03	.01	.15	.07	.03	.01	.13	
L3-SES.female	-.01	.02	-.05	.03	.00	.02	-.04	.04	.00	.02	-.04	.04	.00	.02	-.04	.04	.00	.02	-.04	.04	
L1-MAch	.49	.00	.48	.50	.78	.01	.77	.79	.79	.01	.78	.80	.81	.01	.80	.82	.81	.01	.80	.82	
L1-MAch.female	.06	.00	.05	.07	.06	.01	.04	.08	.06	.01	.04	.08	.06	.01	.04	.08	.06	.01	.04	.08	
L2Mach	-.24	.01	-.25	-.23	-.23	.01	-.24	-.22	-.30	.01	-.32	-.28	-.31	.01	-.33	-.29	-.31	.01	-.33	-.29	H3
L2Mach.female	-.02	.01	-.03	-.01	-.02	.01	-.03	-.01	-.03	.01	-.06	.00	-.03	.01	-.06	.00	-.04	.01	-.07	-.01	
L3Mach	-.18	.03	-.24	-.12	-.19	.03	-.25	-.13	-.18	.03	-.24	-.12	-.19	.03	-.25	-.13	-.19	.03	-.25	-.13	H4
L3Mach.female	-.05	.02	-.09	-.01	-.05	.02	-.09	-.01	-.05	.02	-.09	-.01	-.05	.02	-.09	-.01	-.05	.02	-.09	-.01	
L1-RAch					-.34	.01	-.35	-.33	-.36	.01	-.37	-.35	-.35	.01	-.36	-.34	-.35	.01	-.36	-.34	H5
L1-RAch.female					-.02	.01	-.03	-.01	-.03	.01	-.05	-.01	-.02	.01	-.04	.00	-.02	.01	-.04	.00	
L2-RAch									.05	.01	.04	.06	.05	.01	.04	.06	.05	.01	.04	.06	H6
L2-RAch.female									.01	.01	.00	.02	.01	.01	.00	.02	.01	.01	.00	.02	
Random Effects																					
L3 Country Residual	.03				.03				.03				.03				.02				
Female	.01				.01				.01				.01				.01				
L2 School Residual	.02				.02				.02				.02				.02				
Female	.01				.01				.01				.01				.01				
L1 Student Residual	.77				.75				.75				.74				.74				

Note. Fixed and random effects for multilevel analyses (with three levels: L1: individual student; L2 = school; L3 = country) relating gender, gender equality, socioeconomic status (SES), math achievement, and reading achievement to MSC. The key effects (shaded in grey) are tests of each of the eight a priori hypotheses presented earlier. Values in light blue are not statistically significant (p > .05). All effects are presented as standardized effect sizes (i.e., regression weights in which all first-order effect variables are standardized, M = 0, SD = 1). Est = estimate; SE =standard error; LCI = lower confidence interval; UCI = upper confidence interval;

Hypotheses 3 & 4. Consistent with social comparison theory and the well-established BFLPE (Hypothesis 3), the effect of L2 (school-average) math achievement on MSC is significantly negative (Model 1F; ES =-.24) and remains so in subsequent models. Because this model contains the interaction between gender (coded male = 0, female = 1) this ES = -.24 is the effect for boys. Since the interaction effect (-.03) is small, the effect for girls is only slightly different; -.27 (i.e., -.24 + [-.03 × 1] where -.03 is the interaction and the value of gender is 1 for females; also see Supplemental Table 4 for results excluding gender interactions).

Consistent with the extension of this model to the country level (Hypothesis 4), the effect of country-average (L3) achievement is also significantly negative (Model 1G; ES =-.18). The BFLPE (negative effect of L2-math achievement), remains unchanged by the introduction of L3-math achievement.

Hypotheses 5 & 6. Consistent with DCT (Hypothesis 5), the effect of L1-reading achievement on MSC is significantly negative (Model 1H; ES =-.34) and remains so in subsequent models. Although this effect interacts with gender, the interaction (-.02) is tiny relative to the first-order effect of L1-reading achievement (also see Supplemental Table 4 for models with gender interactions excluded). Consistent with the integration of social and dimensional comparison theories (Hypothesis 6), school-average (L2) reading also has a small, but significantly positive effect (Model 1I; ES = .05) on MSC, that is consistent across subsequent models.

Gender Interactions. Gender effects are moderated by other variables in the analyses, but these interactions are consistently small (all |ESs| < .07). The positive effect of math achievement on MSC was somewhat higher for girls than boys, whereas the negative effect of verbal achievement was somewhat more negative for girls than for boys. Also, negative effects of attending high-ability schools (the BFLPE) and attending high-SES schools were somewhat more negative for girls than boys.

Frame of Reference Effects: Generalizability to Other Outcomes and Gender-Equality Contextual Effects

In testing Hypothesis 7, we applied the final model (Model 1K in Table 1) to additional outcomes frequently evaluated in GEP studies that were a focus of the Else-Quest, et al (2010) study: math self-efficacy, anxiety, instrumental motivation (value), interest (intrinsic motivation), and future plans to pursue mathematics. We also tested the model with additional contextual variables in addition to GGGI (OECD and within-country deviations in achievement, Table 2; but also within-country deviations in HDI, within-country deviations in university enrollment, GGGI and GDI in Supplemental Materials). Although the results in relation to each of the hypotheses are presented in Table 5 (and Supplemental Materials), we highlight important patterns.

Hypothesis 7: Generalizability of Hypothesis 1. Key results (Table 3) are tests of the gender-equality paradox (Hypothesis 1) across the 18 analyses (6 outcomes x 3 gender-parity contextual variables). All but three of these interaction effects are non-significant (consistent with those presented in Appendix 1, considering that the predicted effects are reversed for anxiety that is a negatively oriented outcome). However, there is a small positive interaction for anxiety in relation to GGGI (ES = .05) a result consistent with GEP. In contrast, for the within-country gender deviation in academic achievement, the contextual variable-by-female interaction is significantly negative for anxiety (ES = -.04) and significantly positive for instrumental motivation (ES = .03); both these effects are consistent with a gender stratification hypothesis (and opposite to GEP predictions). In supplemental analyses (see Supplemental Materials), this critical interaction was non-significant for all six outcomes in relation to each of the contextual variables considered (within-country gender differences in HDI and university enrollment, % females in STEM, GDI, and % Muslim). Particularly given the small size and inconsistent direction of the three significant effects, contrasted with large number of non-significant effects, we conclude that the results provide no consistent support for either the GEP or the gender stratification hypothesis.

Table 3A

Generalizability of results to Different Constructs and Alternative Measures of Country-Level Differences in Gender Equality: L3-Contetext = GGGI

	Self-Concept				Self-Efficacy				Anxiety				Utility				Interest				Future Plans				Hypo thesis
Fixed Effects	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	
Female (Fem)	-.06	.01	-.09	-.03	-.14	.01	-.16	-.12	.09	.01	.07	.11	-.03	.01	-.06	.00	*-.01*	.01	-.04	.02	-.12	.01	-.15	-.09	H2
L3-Contetext	*.02*	.03	-.04	.08	*-.01*	.03	-.08	.06	-.13	.03	-.19	-.07	*.07*	.04	-.01	.15	*-.05*	.05	-.14	.04	*.05*	.03	.00	.10	
L3-Context.Fem	*-.02*	.02	-.05	.01	*-.02*	.01	-.04	.00	.05	.01	.02	.08	*.00*	.01	-.03	.03	*.00*	.01	-.03	.03	*-.02*	.02	-.05	.01	H1
L1-SES	.06	.00	.05	.07	.14	.00	.13	.15	-.01	.00	-.02	.00	.04	.00	.03	.05	.03	.00	.02	.04	-.01	.00	-.02	.00	
SES.Fem	-.02	.00	-.03	-.01	-.03	.00	-.04	-.02	*.00*	.00	-.01	.01	-.03	.00	-.04	-.02	-.03	.00	-.04	-.02	-.02	.00	-.03	-.01	
L2-SES	*-.01*	.01	-.02	.00	*-.03*	.01	-.04	-.02	*-.01*	.01	-.02	.00	-.05	.01	-.06	-.04	-.06	.01	-.07	-.05	*.00*	.01	-.01	.01	
L2-SES.Fem	-.03	.01	-.04	-.02	*-.01*	.01	-.02	.00	.02	.01	.01	.03	-.02	.01	-.04	.00	-.02	.01	-.04	.00	*-.01*	.01	-.03	.01	
L3-SES	.07	.03	.01	.13	*.00*	.04	-.08	.08	*-.07*	.03	-.13	-.01	*-.04*	.05	-.13	.05	*-.08*	.05	-.18	.02	*.00*	.03	-.06	.06	
`L3-SES.Fem	*.00*	.02	-.04	.04	-.03	.01	-.05	-.01	*.01*	.02	-.02	.04	*.00*	.02	-.03	.03	*.01*	.02	-.02	.04	*-.03*	.02	-.07	.01	
L1-MAch	.81	.01	.80	.82	.67	.01	.66	.68	-.60	.01	-.61	-.59	.34	.01	.33	.35	.47	.01	.46	.48	.43	.01	.42	.44	
L1-MAch.Fem	.06	.01	.04	.08	-.04	.01	-.06	-.02	-.06	.01	-.08	-.04	.04	.01	.02	.06	.07	.01	.05	.09	*.00*	.01	-.02	.02	
L2Mach	-.31	.01	-.33	-.29	-.07	.01	-.09	-.05	.19	.01	.17	.21	-.09	.01	-.11	-.07	-.14	.01	-.16	-.12	-.09	.01	-.11	-.07	H3
L2Mach.Fem	-.04	.01	-.07	-.01	.03	.01	.01	.05	.04	.01	.01	.07	*-.02*	.01	-.05	.01	-.03	.01	-.06	.00	*-.02*	.01	-.05	.01	
L3Mach	-.19	.03	-.25	-.13	-.09	.04	-.16	-.02	.03	.03	-.03	.09	-.19	.04	-.28	-.10	-.16	.05	-.26	-.06	-.08	.03	-.14	-.02	H4
L3Mach.Fem	-.05	.02	-.09	-.01	*-.01*	.01	-.03	.01	.04	.02	.01	.07	*-.02*	.02	-.05	.01	-.04	.02	-.07	-.01	*.00*	.02	-.03	.03	
L1-RAch	-.35	.01	-.36	-.34	-.17	.01	-.18	-.16	.16	.01	.15	.17	-.13	.01	-.14	-.12	-.26	.01	-.27	-.25	-.30	.01	-.31	-.29	H5
L1-RAch.Fem	-.02	.01	-.04	.00	.02	.01	.00	.04	.06	.01	.04	.08	-.05	.01	-.07	-.03	-.04	.01	-.06	-.02	*-.02*	.01	-.04	.00	
L2-RAch	.05	.01	.04	.06	.02	.01	.01	.03	*-.03*	.01	-.04	-.02	*.01*	.01	.00	.02	.02	.01	.01	.03	.03	.01	.02	.04	H6
L2-RAch.Fem	*.01*	.01	.00	.02	-.01	.01	-.02	.00	-.01	.01	-.02	.00	*.01*	.01	.00	.02	.02	.01	.01	.03	.02	.01	.01	.03	
Random Effects																									
L3 Country	.024				.034				.024				.052				.062				.023				
Fem	.007				.002				.005				.006				.006				.006				
L2 School	.024				.022				.015				.022				.032				.016				
Fem	.006				.000				.003				.004				.004				.003				
L1 Student	.743				.700				.779				.852				.787				.905				

Table 3B

Generalizability of results to Different Constructs and Alternative Measures of Country-Level Differences in Gender Equality: OECD

	Self-Concept				Self-Efficacy				Anxiety				Utility				Interest				Future Plans				Hypo thesis
Fixed Effects	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	
Female (Fem)	-.06	.02	-.10	-.01	-.13	.01	-.16	-.10	.08	.02	.04	.12	-.03	.02	-.06	.01	-.01	.02	-.05	.03	-.11	.02	-.15	-.07	H2
L3-OECD	*-.05*	.06	-.16	.07	*-.08*	.06	-.20	.04	*-.08*	.07	-.21	.05	*-.02*	.08	-.19	.14	-.19	.08	-.35	-.03	*.08*	.06	-.04	.19	
L3-OECD.Fem	*-.02*	*.03*	*-.09*	*.04*	*-.01*	*.02*	*-.06*	*.03*	*.03*	*.03*	*-.02*	*.09*	*-.01*	*.03*	*-.06*	*.05*	*.00*	*.03*	*-.06*	*.06*	*-.02*	*.03*	*-.08*	*.04*	*H1*
L1-SES	.07	.00	.06	.07	.14	.00	.13	.14	-.01	.00	-.02	.00	.04	.00	.04	.05	.03	.00	.02	.03	-.01	.00	-.02	.00	
SES.Fem	-.02	.00	-.03	-.01	-.03	.00	-.04	-.02	-.01	.00	-.01	.00	-.03	.00	-.04	-.02	-.03	.00	-.04	-.02	-.02	.00	-.03	-.01	
L2-SES	-.01	.01	-.02	.00	-.04	.01	-.05	-.03	-.01	.01	-.03	.00	-.05	.01	-.06	-.04	-.07	.01	-.08	-.05	*.00*	.01	-.01	.01	
L2-SES.Fem	-.02	.01	-.04	-.01	-.01	.01	-.02	.01	.02	.01	.00	.03	-.02	.01	-.04	-.01	-.02	.01	-.04	-.01	*-.01*	.01	-.02	.01	
L3-SES	.11	.03	.04	.17	*.02*	.03	-.05	.09	-.13	.03	-.20	-.06	*.02*	.04	-.07	.11	*-.06*	.04	-.15	.02	*.03*	.03	-.03	.09	
`L3-SES.Fem	*.00*	.02	-.03	.04	*-.02*	.01	-.04	.00	*.01*	.02	-.02	.05	*.01*	.02	-.02	.04	*.02*	.02	-.01	.05	*-.03*	.02	-.06	.00	
L1-MAch	.81	.01	.80	.82	.67	.01	.66	.68	-.60	.01	-.61	-.59	.35	.01	.33	.36	.47	.01	.46	.48	.43	.01	.42	.45	
L1-MAch.Fem	.06	.01	.04	.07	-.04	.01	-.06	-.03	-.05	.01	-.07	-.04	.04	.01	.02	.05	.07	.01	.05	.08	.00	.01	-.02	.02	
L2Mach	*-.30*	*.01*	*-.32*	*-.28*	*-.06*	*.01*	*-.08*	*-.04*	*.20*	*.01*	*.18*	*.22*	*-.09*	*.01*	*-.11*	*-.07*	*-.14*	*.01*	*-.16*	*-.12*	*-.09*	*.01*	*-.11*	*-.07*	*H3*
L2Mach.Fem	-.04	.01	-.06	-.01	*.02*	.01	.00	.05	.04	.01	.02	.07	*-.02*	.01	-.04	.01	-.03	.01	-.06	-.01	*-.02*	.01	-.05	.01	
L3Mach	*-.18*	*.03*	*-.24*	*-.13*	*-.09*	*.03*	*-.15*	*-.03*	*.03*	*.03*	*-.03*	*.10*	*-.19*	*.04*	*-.27*	*-.11*	*-.14*	*.04*	*-.22*	*-.06*	*-.10*	*.03*	*-.16*	*-.04*	*H4*
L3Mach.Fem	-.07	.02	-.10	-.03	-.03	.01	-.05	.00	.05	.02	.02	.08	-.03	.01	-.06	-.01	-.06	.01	-.09	-.03	*.00*	.02	-.03	.03	
L1-RAch	*-.35*	*.01*	*-.37*	*-.34*	*-.17*	*.01*	*-.18*	*-.16*	*.17*	*.01*	*.15*	*.18*	*-.13*	*.01*	*-.15*	*-.12*	*-.27*	*.01*	*-.28*	*-.26*	*-.31*	*.01*	*-.32*	*-.30*	*H5*
L1-RAch.Fem	-.02	.01	-.04	-.01	.02	.01	.01	.04	.06	.01	.04	.07	-.04	.01	-.06	-.02	-.04	.01	-.05	-.02	*-.01*	.01	-.03	.00	
L2-RAch	*.05*	*.01*	*.04*	*.06*	*.02*	*.00*	*.01*	*.03*	*-.03*	*.00*	*-.04*	*-.02*	*.01*	*.01*	*.00*	*.02*	*.03*	*.01*	*.02*	*.04*	*.03*	*.01*	*.02*	*.04*	*H6*
L2-RAch.Fem	*.01*	.01	.00	.03	*-.01*	.01	-.02	.00	-.02	.01	-.03	.00	*.01*	.01	-.01	.02	.02	.01	.00	.03	.02	.01	.00	.03	
Random Effects																									
L3 Country	.028				.032				.035				.056				.053				.028				
Fem	.006				.003				.006				.005				.006				.006				
L2 School	.016				.022				.014				.022				.032				.016				
Fem	.003				.000				.003				.004				.004				.003				
L1 Student	.906				.704				.783				.851				.789				.906				

Table 3C

Generalizability of results to Different Constructs and Alternative Measures of Country-Level Differences in Gender Equality: L3Ach-MFDev

	Self-Concept				Self-Efficacy				Anxiety				Utility				Interest				Future Plans				Hypothesis
Fixed Effects	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	Est	SE	LCI	UCI	
Female (Fem)	-.07	.01	-.10	-.04	-.14	.01	-.16	-.12	.10	.01	.08	.12	-.03	.01	-.05	-.01	-.01	.01	-.03	.01	-.12	.01	-.15	-.10	H2
L3Ach-MFDev	-.02	.03	-.07	.03	.01	.03	-.05	.07	.05	.03	-.01	.11	-.02	.04	-.10	.06	.06	.04	-.02	.14	-.03	.03	-.08	.03	
L3Ach-MFDev.Fem	.03	.01	.00	.06	.01	.01	-.01	.03	-.04	.01	-.06	-.02	.03	.01	.01	.05	.02	.01	-.01	.05	.02	.01	.00	.05	H1
L1-SES	.07	.00	.06	.08	.14	.00	.13	.15	-.01	.00	-.02	.00	.04	.00	.03	.05	.03	.00	.02	.04	-.01	.00	-.02	.00	
SES.Fem	-.02	.00	-.03	-.01	-.03	.00	-.04	-.02	-.01	.00	-.02	.00	-.03	.00	-.04	-.02	-.03	.00	-.04	-.02	-.02	.00	-.03	-.01	
L2-SES	-.01	.01	-.02	.00	-.04	.01	-.05	-.03	-.01	.01	-.02	.00	-.05	.01	-.06	-.04	-.07	.01	-.08	-.06	.00	.01	-.01	.01	
L2-SES.Fem	-.02	.01	-.03	-.01	-.01	.01	-.02	.00	.02	.01	.01	.03	-.02	.01	-.04	.00	-.02	.01	-.03	-.01	-.01	.01	-.02	.01	
L3-SES	.10	.03	.04	.16	.00	.03	-.06	.06	-.16	.03	-.23	-.09	.02	.04	-.06	.10	-.11	.04	-.19	-.03	.05	.03	-.01	.11	
`L3-SES.Fem	-.01	.02	-.04	.02	-.02	.01	-.04	.00	.03	.01	.00	.06	.00	.01	-.03	.03	.02	.01	-.01	.05	-.04	.02	-.07	-.01	
L1-MAch	.81	.01	.80	.82	.67	.01	.66	.68	-.60	.01	-.61	-.59	.35	.01	.34	.36	.47	.01	.46	.48	.43	.01	.42	.45	
L1-MAch.Fem	.06	.01	.04	.08	-.04	.01	-.06	-.02	-.05	.01	-.07	-.03	.04	.01	.02	.06	.07	.01	.05	.09	.00	.01	-.02	.02	
L2Mach	-.30	.01	-.32	-.28	-.06	.01	-.08	-.04	.20	.01	.18	.22	-.09	.01	-.11	-.07	-.14	.01	-.16	-.12	-.09	.01	-.11	-.07	H3
L2Mach.Fem	-.04	.01	-.06	-.02	.02	.01	.00	.04	.04	.01	.02	.06	-.02	.01	-.05	.01	-.03	.01	-.06	.00	-.02	.01	-.05	.01	
L3Mach	-.20	.03	-.26	-.14	-.09	.03	-.15	-.03	.04	.03	-.03	.11	-.20	.04	-.28	-.12	-.14	.04	-.22	-.06	-.10	.03	-.16	-.04	H4
L3Mach.Fem	-.06	.02	-.09	-.03	-.03	.01	-.05	-.01	.04	.01	.01	.07	-.02	.01	-.05	.01	-.05	.01	-.08	-.02	.01	.02	-.02	.04	
L1-RAch	-.35	.01	-.36	-.34	-.17	.01	-.18	-.16	.17	.01	.16	.18	-.13	.01	-.14	-.12	-.27	.01	-.28	-.26	-.31	.01	-.32	-.30	H5
L1-RAch.Fem	-.02	.01	-.04	.00	.02	.01	.01	.03	.06	.01	.04	.08	-.04	.01	-.06	-.02	-.04	.01	-.06	-.02	-.01	.01	-.03	.00	
L2-RAch	.05	.01	.04	.06	.02	.00	.01	.03	-.03	.00	-.04	-.02	.01	.01	.00	.02	.03	.01	.02	.04	.03	.01	.02	.04	H6
L2-RAch.Fem	.01	.01	.00	.02	-.01	.01	-.02	.00	-.02	.01	-.03	-.01	.01	.01	.00	.02	.02	.01	.01	.03	.02	.01	.00	.03	
Random Effects																									
L3 Country	.028				.033				.034				.056				.056				.028				
Fem	.007				.003				.005				.005				.006				.006				
L2 School	.024				.022				.014				.022				.032				.016				
Fem	.006				.000				.003				.004				.004				.003				
L1 Student	.745				.704				.783				.851				.789				.906				

Note. Fixed and random effects for multilevel analyses (with three levels: L1: individual student; L2 = school; L3 = country) relating gender, gender equality, socioeconomic status (SES), math achievement, and reading achievement the six self-belief/motivation outcomes. Separate analysis are presented for each of three country-level (L3) measures of gender equality (GGGI, OECD, AchDev; variables are described in greater detail in Appendix 1. The key effects (shaded in grey) are tests of each of the six a priori hypotheses presented earlier. Values in light blue are not statistically significant (p > .05). All effects are presented as standardized effect sizes (i.e., regression weights in which all first-order effect variables are standardized, M = 0, SD = 1). Est = estimate; SE =standard error; LCI = lower confidence interval; UCI = upper confidence interval.

Discussion

The underrepresentation of girls and women in science, technology, engineering, and mathematics (STEM) disciplines is a world-wide phenomenon. This underrepresentation has been stable over many years despite attempts to address the phenomenon. One popular perspective, the gender-stratification hypothesis, proposes that if a country can achieve gender equality, particularly in relation to STEM-related outcomes, then girls will be more likely to pursue STEM subjects. Although there is mixed support for the stratification hypothesis in relation to scores on standardized achievement tests in math and science (Else-Quest et al, 2010), there is little or no support in relation to corresponding levels of self-beliefs and related attitudes in STEM subjects. Indeed, for math self-beliefs there is support for the Gender-Equality Paradox (GEP; Baker & Jones, 1993; Guo, et al., 2019; Marsh, Van Zanden, et al., 2019; Niepel, 2019; Stoet & Geary, 2018; Stoet, Bailey & Moore, 2016), paradoxical findings that countries with greater gender-equality have larger gender gaps disadvantaging girls for MSC, in contrast to gender-stratification hypotheses that gender differences should decline with increases in country-level gender equality and development (Else-Quest, Hyde & Linn, 2010).

So-called GEPs have attracted much attention, despite mixed support and generally small effect sizes. Here, we show that GEPs are very small in size, not robust, and disappear entirely with appropriate control of SES and achievement. Importantly, even without controlling country-SES and achievement, relative indices that juxtapose gender differences within a country (i.e., female/male ratios or female-male differences that hold country-level characteristics constant), show no GEPs and even provide some support for the gender-stratification hypothesis. Therefore, for us, the critical distinction here is not the composite/domain-specific distinction highlighted in this research.

We agree with Else-Quest et al (2010) who claim that composite indices complicate interpretations and can hide important differences associated with specific components that make up the composite, and that specific domains relevant to the outcomes (e.g., country-level differences in achievement for the present investigation) can be more useful. However, in terms of understanding GEPs (and gender stratification), we show that the relative/absolute distinction is more important. The juxtaposition of these two distinctions has important implications for further gender-equality research, but also for the evaluation of existing (and new) country-level measures of gender equality that need to be incorporated into future research.

Although not based in gender equality per se, there are some results that might still be considered paradoxical. In particular, there tend to be larger gender differences in favor of boys in countries where country-levels of math achievement are higher. This is evident for MSC in Model 1G (Table 2), when the interaction between L3-math achievement and gender is first introduced and is unchanged by the introduction of other variables (models 1H – 1K), but also evident for math anxiety and interest (in Table 5). One interpretation is that these effects are too small to be important (|ESs| ≤ .06), even if statistically significant. However, a possible explanation is that girls might have more freedom of choice in more-developed countries, but in less-developed countries girls feel compelled to pursue STEM coursework in support of limited economic opportunities (Else-Quest et al., 2010; also see Stoet & Geary, 2018; in less-developed countries there may be pressure on girls from families to pursue STEM as a safe option in terms of career choice. Relatedly, Charles and Bradley (2009) found girls' affinity for math was higher in developing countries than advanced industrial countries, suggesting that girls are more likely to pursue and have affinity for math when getting a stable, well-paying job is a priority.

We speculate on another related possibility based on results by Marsh, Abduljabbar et al. (2014) comparing gender differences in the US and Saudi Arabia as well as our finding in the present paper that gender differences are significantly smaller in countries with more Muslim students. Marsh, Abduljabbar et al. noted that among countries where girls out-perform boys in math and science, nearly all are Middle Eastern Islamic countries—even though achievement levels overall tend to be low in these countries. To juxtapose this finding with GEPs, we included the percent of Muslims as one of our country-level contextual variables. In countries with larger Muslim populations, girls had higher MSCs than boys (Table 1), a finding that might be interpreted as support for GEPs. However, Marsh, Abduljabbar et al. noted that in Saudi Arabia, girls tend to spend more time and effort in schoolwork than boys. Indeed, Saudi boys tend to have more resources and freedom in how to spend their time, whereas school is one of a more restricted range of activities in which girls are free to excel. In this respect, the results are consistent with the suggestion that girls in less gender-equal countries are likely to invest more time and effort in schoolwork, resulting in better academic outcomes. Furthermore, as emphasized by Marsh, Abduljabbar et al., school systems in these countries are more likely to be single-sex schools. Hence the frame-of-reference used by girls is based almost exclusively on the performances of other girls, rather than boys. Although beyond the scope of the present investigation, we note the need for more contextualized studies of cross-national studies of gender differences in STEM achievement and self-beliefs.

Our study draws heavily on recent advances in ASC theory, particularly in relation to the seemingly paradoxical frame-of-reference effects in social and dimensional comparison theories that have been largely absent in GEP research. The integration of these theoretical perspectives with cross-national gender-equality research provides new explanations for paradoxical findings—particularly GEPs, but also how domain-specific self-perceptions in one domain can be influenced by accomplishments in an entirely different domain. The gender difference in MSC is only marginally reduced by controlling for math achievements. Yet, paradoxically, these gender differences are substantially reduced by controlling for verbal achievements. Hence, girls' lower MSCs are apparently more a function their much higher (relative to boys) verbal skills than their lower math skills. This seemingly paradoxical finding is a straight-forward extension of the well-established DCT. Indeed, controlling for both math and verbal achievements nearly eliminated gender differences in MSC (ES = -.06), an effect so small that it can be considered to be close to zero and trivial in size. Although this joint reliance on multiple domains of achievement complicates results, it also opens fascinating new areas of research on the processes underlying these effects—both in terms of prediction, but also in terms of long-term implications (e.g., coursework selection and career choice) and intervention studies that target one domain in isolation of other domains (Else-Quest & Hamilton, 2018).

In support of the cross-cultural generalizability of the results, it is relevant to juxtapose the large-scale cross-cultural research used here with traditional meta-analyses approaches typically used to evaluate gender differences. Increasingly, both use evolving multi-level analyses that focus on both overall ESs and residual variance (study-to-study or country-to-country variation). Meta-analysis, despite its many strengths, typically suffers from the heterogeneity of studies in respect to materials, participants, measures, research designs, publication bias, an over-reliance on middle-class and Western participants, and an over-representation of studies published in English-language journals. Importantly, because meta-analysts rarely have access to individual-level data, appropriate multilevel tests (e.g., effects of L1, L2, and L3 math achievement on MSC) are not possible. Large-scale cross-national studies like the present investigation overcome many of the problems and provide a stronger basis for evaluating the universality of findings, but still have potential weaknesses in relation to generalizability. Else-Quest, et al. (2010) similarly argued for the advantages of large cross-national studies but used a traditional mixed-effects meta-analysis that did not fully exploit the multilevel structure of their PISA2003 data. There are, of course, limitations in the use of large-scale secondary databases like PISA as well as strengths. In particular, PISA is based on responses by 15-year-olds to a self-selected sample of countries using a single set of measures of each construct that might not generalize to results based on different measures of the same constructs. Importantly, meta-analyses and large cross-national databases are not mutually exclusive such that juxtaposing the results of both approaches within the same study provides stronger tests of generalizability/universality of the findings (see Möller et al., 2011). Future substantive-methodological synergies that integrate cutting-edge theoretical development, design, and statistical analyses, will provide further insight into these complex issues.

References

Baker, D. P., & Jones, D. P. (1993). Creating gender equality: Cross-national gender stratification and mathematical performance. *Sociology of Education, 66,* 91–103. http://dx.doi.org/10.2307/2112795

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods, 16(3),* 265-284. http://dx.doi.org/1.1037/a0024448

Ceci, S. J., Ginther, D. K., & Williams, W. M. (2014). Women in Academic Science. *Psychological Science in the Public Interest, 15(3),* 75-141. https://doi.org/10.1126/science.266.5189.1327-b

Charles, M. & Bradley, K. (2009). Indulging our gendered selves? Sex segregation by field of study in 44 countries. American Journal of Sociology, 114, 924-976. https://doi.org/10.1086/595942

Cohen, P., West, S. G., & Aiken, L. S. (2014). *Applied multiple regression/correlation analysis for the behavioral sciences.* Psychology Press

Eccles, J. S. (2009). Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist, 44,* 78–89. https://doi.org/10.1080/00461520902832368

Else-Quest, N. M., & Hamilton, V. (2018). *Measurement and analysis of nation-level gender equality in the psychology of women.* In C. B. Travis & J. W. White (Eds.), Handbook of the psychology of women: Vol. 2. Perspectives on women’s private and public lives (pp. 545–563). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/0000060-029

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*, 103. doi:10.1037/a0018053

Guo, J., Marsh, H. W., Parker, P. D., Dicke, T. (2019). Cross-cultural generalizability of social and dimensional comparison effects on reading, math, and science self-concepts for primary school students using the combined PIRLS and TIMSS data. *Learning and Instruction, 58,* 210-219. doi.org/10.1016/j.learninstruc.2018.07.007

Guo, J., Marsh, H. W., Parker, P. D., Dicke, T., Van Zanden, B. (2018). Countries, parental occupation, and girls' interest in science. *Lancet, 393,* e6-e8. doi: https://doi.org/10.1016/s0140-6736(19)30210-7

Guo, Marsh, H. W., Parker, P. D., Morin, A.J.S., & Dicke, T. (2017). Extending expectancy-value theory predictions of achievement and aspirations in science: Dimensional comparison processes and expectancy-by-value interactions. *Learning and Instruction, 49,* 81-91. https://doi.org/10.1016/j.learninstruc.2016.12.007

Hawken, A., & Munck, G. L. (2013). Cross-national indices with gender-differentiated data: what do they measure? How valid are they?. *Social Indicators Research*, *111*(3), 801-838. doi: https://doi.org/10.1007/s11205-012-0035-7

Hofstede, G.(1984). *Culture's Consequences: International Differences in Work-Related Values* (2nd ed.). Beverly Hills CA: SAGE Publications. ISBN 0-8039-1444-X.

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

Hyde, J. S. (2007). New directions in the study of gender similarities and differences. *Current Directions in Psychological Science, 16*, 259-263. doi:10.1111/j.1467-8721.2007.00516.x

Hyde, J. S. (2012). Nation-level indicators of gender equality in psychological research: Theoretical and methodological issues. *Psychology of Women Quarterly, 36*, 145-148. https://doi.org/10.1177/0361684312441448

James, W. (1890/1963). *The principles of psychology* (Vol. 2). New York: Holt.

Klasen, S. (2018). The Impact of Gender Inequality on Economic Performance in Developing Countries. *Annual Review of Resource Economics, 10,* 279-298. https://doi.org/10.1146/annurev-resource-100517-023429

Marsh, H. W. (1984). Self-concept: The application of a frame of reference model to explain paradoxical results. *Australian Journal of Education, 28*, 165–181. https://doi.org/10.1177/000494418402800207

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23*(1), 129-149. Retrieved from http://journals.sagepub.com/doi/pdf/10.3102/00028312023001129

Marsh, H. W. (1989). Age and sex effects in multiple dimensions of self-concept: Preadolescence to Early-adulthood. *Journal of Educational Psychology, 81*, 417-430. https://doi.org/10.1037/0022-0663.81.3.417

Marsh, H. W. (1994). Using the National Educational Longitudinal Study of 1988 to evaluate theoretical models of self-concept: The Self-Description Questionnaire. *Journal of Educational Psychology, 86,* 439–456. https://doi.org/10.1037/0022-0663.86.3.439

Marsh, H. W. (2007). *Self-concept theory, measurement and research into practice: The role of self-concept in educational psychology*. London, England: British Psychological Society.

Marsh, H. W. (2016). Cross-cultural generalizability of year in school effects: Negative effects of acceleration and positive effects of retention on academic self-concept. *Journal of Educational Psychology, 108*(2), 256–273. http://dx.doi.org/10.1037/edu0000059

Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J. S., Abdelfattah, F., & Nagengast, B. (2014). The Big-Fish-Little-Pond effect in mathematics: A cross-cultural comparison of U.S. and Saudi Arabian TIMSS responses. *Journal of Cross-Cultural Psychology*, 45(5), 777-804. doi:10.1177/0022022113519858

Marsh, H. W. & Parker, J. W. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don’t learn to swim as well? *Journal of Personality and Social Psychology, 47*(1), 213–231. https://doi.org/10.1037/0022-3514.47.1.213

Marsh, H. W., Parker, P. & Craven, R. G. (2015). Dimensional Comparisons Theory: An Extension of the Internal/External Frame of Reference Model. In F. Guay, F., H. W. Marsh, R. G. Craven, & D. McInerney, D. (Eds.). (2015). *Self-concept, Motivation, and Identity: Underpinning success with research and practice.* International Advances in Self Research (Vol 5), pp. 115–151. Charlotte, NC: Information Age Publishing.

Marsh, H. W., Pekrun, R., Murayama, K., Arens, A. K., Parker, P. D., Guo, J., & Dicke, T. (2018). An Integrated Model of Academic Self-Concept Development: Academic Self-Concept, Grades, Test Scores, and Tracking Over 6 Years. *Developmental Psychology, 54,* 263-280. http://dx.doi.org/10.1037/dev0000393

Marsh, H. W., & Seaton, M. (2015). The Big-fish–little-pond effect, competence self-perceptions, and relativity: Substantive advances and methodological innovation. In A. J. Elliott (Ed.). *Advances in Motivation Science,* 2, 127–184). New York Elsevier.

Marsh, H. W., Zanden, B. V., Parker, P. D., Guo, J., Conigrave, J., & Seaton, M. (2019). Young Women Face Disadvantage to Enrollment in University STEM Coursework Regardless of Prior Achievement and Attitudes. *American Educational Research Journal.* Advanced online publication. doi:10.3102/0002831218824111

Möller, J., & Marsh, H. W. (2013). Dimensional comparison theory. *Psychological Review*, *120*(3), 544. doi:10.1037/a0032459

Möller, J., Retelsdorf, J., Köller, O., Marsh H. W. (2011). The reciprocal I/E model: An integration of models of relations between academic achievement and self-concept. *American Educational Research Journal, 48,* 1315-1346. doi:10.1016/b978-0-08-044894-7.00619-9

Niepel, C., Stadler, M., & Greiff, S. (2019). Seeing is believing: Gender diversity in STEM is related to mathematics self-concept. *Journal of Educational Psychology, 111(6),* 1119-1130. https://doi.org/10.1037/edu0000340

OECD. (2014). *PISA 2012 Technical Report.* Paris: OECD

Parker, P. D., Marsh, H. W., Lüdtke, O., & Trautwein, U. (2013). Differential school contextual effects for math and English: Integrating the big-fish-little-pond effect and the internal/external frame of reference. *Learning and Instruction, 23,* 78–89. doi:10.1016/j.learninstruc.2012.07.001

Rasbash, J., Steele, F., Browne, W. & Prosser, B. (2004). *A user's guide to MLwiN–Version 2.0.* University of Bristol.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Stoet, G., Bailey, D. H., Moore, A. M., & Geary, D. C. (2016). Countries with higher levels of gender equality show larger national sex differences in mathematics anxiety and relatively lower parental mathematics valuation for girls. *PloS one*, *11*, e0153857. doi:10.1371/journal.pone.0153857

Stoet, G. & Geary, D. (2013). Sex Differences in Mathematics and Reading Achievement Are Inversely Related: Within- and Across-Nation Assessment of 10 Years of PISA Data. *PloS One., 8(3),* E57988. https://doi.org/10.1371/journal.pone.0057988

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

Stoet G, Geary DC (2019) A simplified approach to measuring national gender inequality. PLoS ONE 14(1): e0205349. https://doi.org/10.1371/journal.pone.0205349

Wang, M.-T., & Degol, J. L. (2017). Who chooses stem careers? Using a relative cognitive strength and interest model to predict careers in science, technology, engineering, and mathematics. *Journal of Youth Adolescence, 46,* 1805–182. DOI 1.1007/s10964-016-0618-8

Wang, M. T., Eccles, J., & Kenny, S. (2013). Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics. *Psychological Science, 24*(5), 770-775. doi: 10.1177/0956797612458937

Wylie, R. C. (1979). *The self-concept (Vol. 2).* Lincoln, NE: University of Nebraska Press

Appendix 1

Key Constructs Used in the Present Investigation.

Country-Level Contextual Variables

Human Development Index (HDI) composite index of life expectancy, education, and per capita income indicators. In addition to this absolute measure, we included separate HDI scores for men and women, female/male ratios, and female-male differences.

Socioeconomic status (SES). Individual SES is a PISA2012 composite index of economic, social and cultural status based on highest occupational status of parents, highest educational level of parents, and home possessions (including educational resources and number of books). SES was then aggregated up to the school and country levels with responses standardized (mean = 0 and SD =1) separately at each level.

OECD. Organisation for Economic Co-operation and Development Country (1=yes, 0=n0)

Gender-Related Development Index (GDI) is the ratio of HDIs components (health, knowledge and living standards) computed as the female/male ratio (see Klasen, 2006, who suggested that a better measure would be computation of HDI separately for males and females for each country and the comparison of these values).

Global Empowerment Measure (GEM) is a composite measure to quantify the loss of achievement within a country due to gender inequality based on reproductive health, empowerment, and labor market participation (but excludes education)

Tertiary Enrollment was used to construct ratio and difference scores based on the percent of the female population of official school age enrolled in tertiary education and the percent of the male population of official school age enrolled in tertiary education

Global Gender Gap Index (GGGI) is a composite measure of female/male ratios for a diverse set of absolute and relative measures based on participation and opportunity, educational attainment, political empowerment, and health/survival. Ratios for individual components are truncated to 1.0 if females score higher than males.

Gross Domestic Product (GDP) is a gross is a monetary measure of the market value of all the final goods and services

Achievement Country-average achievement measures for math, science, reading and their average were based on PISA results. For total achievement we also considered female-male differences within each country as a relative measure.

%Female-STEM (%F-STEM). Is the female share of graduated in science, math, engineering, manufacturing and construction at the tertiary level,

%Female-STEM (%F-STEM %F-Manage Female share of employment in senior and middle level management, standardized (M = 0, SD =1) across all countries

Human Development Index-loss (HDI-loss) is the United Nations measures of loss of human development due to inequality

Gender Inequality Index (GII) is based on three components of human development—reproductive health, empowerment, and economic status (high values reflect greater inequality)

Power distance index. Hofstede's.(1984) measure in which high scores reflect a strong hierarchical structure and low scores indicate power is more widely distributed.

Individualism/collectivism is Hofstede's.(1984) measure in which high scores reflect societies with loose ties between individuals with family and groups, and low scores reflect strongly integrated individual ties with extended families and groups

Individual Student (L1) Measures of Self-Belief & Motivation

Math Self-concept. (MSC) In my mathematics class, I understand even the most difficult work

Math Self-efficacy. How confident are you about solving a problem like Solving an equation like 2(x+3) = (x + 3) (x - 3)

Math Anxiety. I often worry that it will be difficult for me in mathematics classes

Math instrumental motivation: Making an effort in mathematics is worth it because it will help me in the work that I want to do later on

Math intrinsic motivation: I look forward to my mathematics lessons

Math Future Intentions: Paired-comparison items pitting math intentions against other subjects. Choose one of the following: "I intend to take additional mathematics courses after school finishes" or "I intend to take additional <test language> courses after school finishes"

Note. For present purposes all scores were standardized (Mn = 0, SD = 1) separately at each level. Thus, for example, individual student achievement and SES scores were separately standardized at the individual-student, school-average, and country levels. Other country-level measures were standardized at the country level. GEP tests were based on the interaction between female gender (male = 0, female = 1) and standardized country-level variables.

Supplemental Materials

Supplemental Table 1. *Countries: Number of Students and Number of Schools*

Supplemental Table 2. *Variance Components for Selected Variables*

Supplemental Table 3. *Gender Equality Paradox: Fixed and Random Effects for Selected Variables*

Supplemental Table 4.
Math Self-concept: Effects of Gender, Gender Equality (GGGI), SES, Math achievement, and Reading Achievement: Excluding Gender Interactions and/or Covariates

Supplemental Table 5. *Generalizability of results to Different Constructs and Alternative Measures of Country-Level Differences in Gender Equality*

Supplemental Material 6. Computer code used in the analysis of the data (macros for the MLwiN statistical package)

Supplemental Table 1

Countries: Number of Students and Number of Schools

country	Number of students	Number of schools
Albania	4743	204
United Arab Emirates	11500	458
Argentina	5908	226
Australia	14481	775
Austria	4755	191
Belgium	8597	287
Bulgaria	5282	188
Brazil	19204	839
Canada	21544	885
Switzerland	11229	411
Chile	6856	221
Colombia	9073	352
Costa Rica	4602	193
Czech Republic	5327	297
Germany	5001	230
Denmark	7481	341
Spain	25313	902
Estonia	4779	206
Finland	8829	311
France	4613	226
United Kingdom	12659	507
Greece	5125	188
Hong Kong-China	4670	148
Croatia	5008	163
Hungary	4810	204
Indonesia	5622	209
Ireland	5016	183
Iceland	3508	134
Israel	5055	172
Italy	31073	1194
Jordan	7038	233
Japan	6351	191
Kazakhstan	5808	218
Korea	5033	156
Liechtenstein	293	12
Lithuania	4618	216
Luxembourg	5258	42
Latvia	4306	211
Macao-China	5335	45
Mexico	33806	1471
Montenegro	4744	51

Malaysia	5197	164
Netherlands	4460	179
Norway	4686	197
New Zealand	4291	177
Peru	6035	240
Poland	4607	184
Portugal	5722	195
Qatar	10966	157
Shanghai-China	5177	155
Perm(Russian Federation)	1761	63
Florida (USA)	1896	54
Connecticut (USA)	1697	50
Massachusetts (USA)	1723	49
Romania	5074	178
Russian Federation	5231	227
Singapore	5546	172
Serbia	4684	153
Slovak Republic	4678	231
Slovenia	5911	338
Sweden	4736	209
Chinese Taipei	6046	163
Thailand	6606	239
Tunisia	4407	153
Turkey	4848	170
Uruguay	5315	180
United States of America	4978	162
Viet Nam	4959	162
Total	485490	18292

Note. Included here are the 68 countries/territories that are the basis of the present investigation. Also included are the number of schools and the number of students from each of the countries. For each country there a two-stage sampling design (random selection of schools and then random selection of students within each school) to achieve a nationally representative sample of students from each country.

Supplemental Table 2

Variance Components for Selected Variables

	Variance Components Due to:		
Responses	Country	School	Student
Achievement Scores			
Achievement Scores			
Math	.260	.289	.466
Science	.228	.290	.496
Reading	.201	.330	.504
Math Self-Belief/Motivation Constructs			
Math Attitudes			
Self-concept	.039	.029	.924
Self-efficacy:	.056	.087	.864
Anxiety	.066	.027	.905
Instrumental	.081	.026	.890
Interest	.116	.036	.846
Intentions	.029	.022	.955

Note. Variance components are based on a three-level model (L1: individual student; L2 = school; L3 = country) with no predictor variables. The results indicate that there is substantial variance variability for achievement variables at the level of school and country, as well as at the level of individual student. In contrast, for the math self-belief/motivation constructs, most of the variance is at the level of the individual student. This is, of course, consistent with the frame-of-reference models and ASC theory, positing that students' self-beliefs are primarily a function of how they compare with students within their school (and country). However, because there is so much variance in achievement measures at the school and country levels, it means that school and country levels of achievement (as well as individual student achievement) will be related to self-beliefs.

Supplemental Table 3

Gender Equality Paradox: Fixed and Random Effects For Selected Variables (also see Table 3 in main text)

L3 Contextual Variable	Fixed Effects			Random Effects				
	Female	L3-Var	L3var.Female (Paradox)	L3-Res	Female	L2-Resid	Female	L1-Resid
No Contextual Variables (68)	-.30			.038	.018	.027	.005	.898
Human Development Index (HDI, 60)	-.29	.04	-.09	.031	.011	.026	.006	.896
Gender Inequality Index (53)	-.28	.03	-.06	.032	.014	.027	.005	.895
Global Equality Measure (59)	-.29	.04	-.08	.031	.011	.027	.006	.898
OECD (68)	-.24	.00	-.12	.037	.014	.027	.005	.898
Gross Domestic Product (63)	-.29	.06	-.07	.031	.013	.027	.005	.898
SES (68)	-.29	.07	-.05	.033	.015	.027	.005	.898
Math Achievement (68)	-.29	-.03	-.07	.036	.013	.027	.005	.898
Reading Achievement (68)	-.29	-.03	-.08	.037	.012	.027	.005	.898
Science Achievement (68)	-.29	-.03	-.07	.037	.013	.027	.005	.898
Total Achievement (68)	-.29	-.03	-.08	.036	.012	.027	.005	.898
Human Development Index (HDI, 60)	-.29	.04	-.09	.031	.011	.026	.006	.896
HDI-Income loss (56)	-.29	03	-.08	.030	.011	.026	.005	.902
Global Gender Gap Index (55)	-.28	.03	-.06	.031	.014	.026	.006	.900
HDI M-F Deviation (60)	-.29	.14	.01	.032	.018	.026	.006	.897
HDI-M/F-Ratio (60)	-.29	-.02	-.01	.032	.018	.026	.006	.897
Gender Development Index (60)	-.29	-.01	-.01	.032	.018	.026	.006	.897
Tertiary Enrollment F-M Deviation	-.29	.04	.01	.030	.017	.026	.006	.894
Tertiary Enrollment F/M ratio	-.29	.06	.03	.030	.017	.026	.006	.894
Total Achievement F-M Deviation (68)	-.30	.01	.06	.037	.015	.027	.005	.898

Note. Separate three-level analyses were done for each contextual variable, predicting individual math self-concept with individual gender (female), the country-level contextual variable, and their interaction. Fixed and random effects for multilevel analyses (with three levels: L1: individual student; L2 = school; L3 = country) relating gender, each done with each of L3-Contextual variables, their interaction to math self-concept. The key effect are tests of the gender-equality paradox (supported by a significantly negative L3-contextual variable-by-female interaction). Numbers in parentheses are the numbers of countries/regions for which information was available and included in the analyses.

Supplemental Table 4

Math Self-concept: Effects of Gender, Gender Equality (GGGI), SES, Math achievement, and Reading Achievement: Excluding Gender Interactions and/or Covariates

	Excluding Interactions and covariates				Excluding Only Interactions					Hypo-thesis
	M2b	M2c	M2d	M2e	M3a	M3b	M3c	M3d	M3e	
Fixed Effects										
L3-GGGI					.03	.00	.02	.02	.02	
female	-.23	-.23	-.08	-.08	-.29	-.22	-.22	-.07	-.07	
L1-SES						.04	.04	.05	.05	
L2-SES						-.04	-.04	-.02	-.03	
L3-SES						.02	.10	.08	.09	
L1-MAch	.52	.52	.82	.83		.51	.51	.81	.83	
L2Mach	-.26	-.26	-.23	-.32		-.25	-.25	-.24	-.32	H3
L3Mach		-.12	-.14	-.13			-.18	-.19	-.19	H4
L1-RAch			-.35	-.37				-.35	-.37	H5
L2-RAch				.06					.05	H6
Random Effects										
L3 Country Resid	.05	.04	.04	.04	.03	.05	.03	.03	.03	
Female	.01	.01	.01	.01	.02	.01	.01	.01	.01	
L2 School Resid	.02	.02	.02	.02	.03	.02	.02	.02	.02	
Female	.01	.01	.01	.01	.01	.01	.01	.01	.01	
L1 Student Resid	.77	.77	.75	.75	.90	.77	.77	.75	.75	

Note. Fixed and random effects for multilevel analyses (with three levels: L1: individual student; L2 = school; L3 = country) relating gender, gender equality, socioeconomic status (SES), math achievement, and reading achievement to MSC. Models 2b-2e exclude gender interactions and covariaites. Models 3a-3e exclude only gender interactions. Models including both covariates and gender interactions are presented in Table 2 of the main text. The key effects (shaded in grey) are test of each of the a priori hypotheses presented earlier. Values in light gray are not statistically significant (p > .05). All effects are presented as standardized effect sizes (i.e., regression weights in which all first-order effect variables are standardized, M = 0, SD = 1)

Supplemental Table 5A

Generalizability of results to Different Constructs and Alternative Measures of Country-Level Differences in Gender Equality (also see Table 5 in Main text)

		L3-Context = GGGI						L3-Context = OECD						L3-Context = L3Ach-MFDev					
Fixed Effects		SC	SE	Anx	Util	Int	FP	SC	SE	Anx	Util	Int	FP	SC	SE	Anx	Util	Int	FP
Female (Fem)		-.06	-.14	.09	-.03	*-.01*	-.12	-.06	-.13	.08	*-.03*	*-.01*	-.11	-.07	-.14	.10	-.03	-.01	-.12
L3-Context		*.02*	*-.01*	-.13	*.07*	*-.05*	*.05*	*-.05*	*-.08*	*-.08*	*-.02*	-.19	*.08*	-.02	*.01*	*.05*	*-.02*	*.06*	*-.03*
L3-Context.Fem	H1	*-.02*	*-.02*	.05	*.00*	*.00*	*-.02*	*-.02*	*-.01*	*.03*	*-.01*	*.00*	*-.02*	.03	*.01*	-.04	.03	*.02*	*.02*
L1-SES		.06	.14	-.01	.04	.03	-.01	.07	.14	-.01	.04	.03	-.01	.07	.14	-.01	.04	.03	-.01
SES.Fem		-.02	-.03	*.00*	-.03	-.03	-.02	-.02	-.03	-.01	-.03	-.03	-.02	-.02	-.03	*-.01*	-.03	-.03	-.02
L2-SES		*-.01*	*-.03*	*-.01*	-.05	-.06	*.00*	-.01	-.04	-.01	-.05	-.07	*.00*	-.01	-.04	-.01	-.05	-.07	*.00*
L2-SES.Fem		-.03	*-.01*	.02	-.02	-.02	*-.01*	-.02	-.01	.02	-.02	-.02	*-.01*	-.02	*-.01*	.02	-.02	-.02	*-.01*
L3-SES		.07	*.00*	*-.07*	*-.04*	*-.08*	*.00*	.11	*.02*	-.13	*.02*	*-.06*	*.03*	.10	*.00*	-.16	*.02*	-.11	*.05*
`L3-SES.Fem		*.00*	-.03	*.01*	*.00*	*.01*	*-.03*	*.00*	*-.02*	*.01*	*.01*	*.02*	*-.03*	*-.01*	-.02	.03	*.00*	*.02*	-.04
L1-MAch		.81	.67	-.60	.34	.47	.43	.81	.67	-.60	.35	.47	.43	.81	.67	-.60	.35	.47	.43
L1-MAch.Fem		.06	-.04	-.06	.04	.07	*.00*	.06	-.04	-.05	.04	.07	.00	.06	-.04	-.05	.04	.07	*.00*
L2Mach	H3	-.31	-.07	.19	-.09	-.14	-.09	-.30	-.06	.20	-.09	-.14	-.09	-.30	-.06	.20	-.09	-.14	-.09
L2Mach.Fem		-.04	.03	.04	*-.02*	-.03	*-.02*	-.04	*.02*	.04	*-.02*	-.03	*-.02*	-.04	.02	.04	*-.02*	-.03	*-.02*
L3Mach	H4	-.19	-.09	.03	-.19	-.16	-.08	-.18	-.09	*.03*	-.19	-.14	-.10	-.20	-.09	*.04*	-.20	-.14	-.10
L3Mach.Fem		-.05	*-.01*	.04	*-.02*	-.04	*.00*	-.07	-.03	.05	-.03	-.06	*.00*	-.06	-.03	.04	*-.02*	-.05	*.01*
L1-RAch	H5	-.35	-.17	.16	-.13	-.26	-.30	-.35	-.17	.17	-.13	-.27	-.31	-.35	-.17	.17	-.13	-.27	-.31
L1-RAch.Fem		-.02	.02	.06	-.05	-.04	*-.02*	-.02	.02	.06	-.04	-.04	*-.01*	-.02	.02	.06	-.04	-.04	*-.01*
L2-RAch	H6	.05	.02	*-.03*	*.01*	.02	.03	.05	.02	-.03	*.01*	.03	.03	.05	.02	-.03	*.01*	.03	.03
L2-RAch.Fem		*.01*	-.01	-.01	*.01*	.02	.02	*.01*	*-.01*	-.02	*.01*	.02	.02	*.01*	*-.01*	-.02	*.01*	.02	*.02*
ZGRADE	H7	-.06	-.03	.03	-.06	-.06	*.00*	-.06	-.03	.03	-.06	-.06	-.01	-.06	-.03	.03	-.06	-.06	*-.01*
ZGRADE.Fem		*.00*	*.00*	*.01*	*-.01*	*.00*	*-.01*	*.00*	*.00*	*.01*	*-.01*	*.00*	*-.01*	*.00*	*.00*	*.01*	*-.01*	*.00*	*-.01*
ImmG1	H8A	.16	.14	-.07	.21	.27	.05	.15	.13	-.06	.20	.26	.04	.15	.13	-.06	.20	.26	.04
ImmG1.Fem		*.01*	*.03*	*-.01*	*-.02*	*-.01*	*.01*	*.01*	.03	*-.01*	*-.01*	*.00*	*.01*	*.01*	.03	*-.01*	*-.01*	*.00*	*.01*
ImmG2	H8B	.10	.12	.01	.16	.17	.08	.09	.11	*.01*	.15	.16	.08	.09	.11	.01	.15	.16	.08
ImmG2.Fem		*.01*	*.00*	*-.04*	*-.04*	*-.03*	*.00*	*.00*	*.00*	*-.03*	-.04	-.03	*-.01*	*.00*	*.00*	*-.03*	-.04	*-.03*	*-.01*
Random Effects																			
L3 Cntry		.024	.034	.024	.052	.062	.023	.028	.032	.035	.056	.053	.028	.028	.033	.034	.056	.056	.028
Fem		.007	.002	.005	.006	.006	.006	.008	.003	.006	.005	.006	.006	.007	.003	.005	.005	.006	.006
L2 School		.024	.022	.015	.022	.032	.016	.024	.022	.014	.022	.032	.016	.024	.022	.014	.022	.032	.016
Fem		.006	.000	.003	.004	.004	.003	.006	.000	.003	.004	.004	.003	.006	.000	.003	.004	.004	.003
L1 Student		.743	.700	.779	.852	.787	.905	.745	.704	.783	.851	.789	.906	.745	.704	.783	.851	.789	.906

Table 5B

Generalizability of results to Different Constructs and Alternative Measures of Country-Level Differences in Gender Equality

		L3-Context = HDIdiff						L3-Context =MF Difference University Enrollment						L3-Context = %F-STEM					
Fixed Effects		SC	SE	Anx	Util	Int	FP	SC	SE	Anx	Util	Int	FP	SC	SE	Anx	Util	Int	FP
Female (Fem)		-.06	-.14	.09	-.03	-.01	-.12	-.06	-.14	.09	-.03	.00	-.13	-.06	-.14	.09	-.02	.00	-.12
L3-Context		.03	.01	.02	-.01	.02	-.01	-.01	.02	-.03	.03	-.04	.02	-.02	-.01	.06	-.01	.03	-.04
L3-Context.Fem	H1	-.02	-.01	.01	.00	-.02	.01	.01	.00	.00	.02	.02	-.02	.03	.01	-.02	.02	.02	.01
L1-SES		.07	.14	-.01	.04	.03	-.01	.06	.13	-.01	.04	.02	-.01	.07	.13	-.01	.04	.03	-.01
SES.Fem		-.02	-.03	-.01	-.03	-.03	-.02	-.02	-.03	.00	-.02	-.03	-.02	-.02	-.03	-.01	-.03	-.03	-.02
L2-SES		-.01	-.04	-.01	-.05	-.07	.00	-.01	-.04	-.01	-.05	-.07	.00	-.01	-.03	-.02	-.05	-.06	.00
L2-SES.Fem		-.02	-.01	.02	-.02	-.02	-.01	-.03	-.01	.02	-.03	-.02	-.01	-.02	-.01	.02	-.02	-.02	-.01
L3-SES		.08	.01	-.12	.00	-.09	.03	.12	.00	-.10	.03	.04	.04	.08	.01	-.09	.01	-.05	.01
`L3-SES.Fem		-.01	-.03	.03	.00	.01	-.03	-.02	-.03	.02	-.01	-.02	-.02	-.02	-.03	.03	.00	.00	-.03
L1-MAch		.81	.66	-.60	.35	.47	.43	.80	.66	-.59	.34	.46	.42	.80	.66	-.59	.34	.46	.42
L1-MAch.Fem		.06	-.04	-.06	.04	.07	.00	.07	-.04	-.06	.04	.08	.00	.07	-.04	-.06	.04	.08	.00
L2Mach	H3	-.31	-.07	.19	-.09	-.14	-.09	-.30	-.07	.19	-.08	-.14	-.08	-.30	-.07	.18	-.08	-.13	-.07
L2Mach.Fem		-.04	.03	.04	-.02	-.03	-.02	-.04	.02	.04	-.03	-.03	-.02	-.04	.03	.05	-.02	-.03	-.02
L3Mach	H4	-.19	-.13	.01	-.19	-.19	-.08	-.24	-.14	-.03	-.23	-.33	-.08	-.22	-.16	-.03	-.22	-.28	-.08
L3Mach.Fem		-.06	-.03	.05	-.03	-.04	-.01	-.04	-.02	.05	-.01	-.01	-.01	-.04	-.02	.04	-.01	-.02	-.01
L1-RAch	H5	-.35	-.17	.16	-.13	-.26	-.30	-.35	-.17	.16	-.13	-.26	-.30	-.34	-.17	.16	-.13	-.26	-.29
L1-RAch.Fem		-.02	.02	.06	-.04	-.04	-.01	-.02	.02	.06	-.05	-.04	-.02	-.02	.02	.06	-.05	-.04	-.02
L2-RAch	H6	.05	.02	-.03	.01	.03	.03	.05	.02	-.03	.01	.02	.03	.05	.02	-.03	.01	.03	.03
L2-RAch.Fem		.01	-.01	-.02	.01	.02	.02	.01	-.01	-.02	.01	.02	.02	.01	-.01	-.02	.01	.02	.02
ZGRADE		-.06	-.03	.03	-.06	-.06	.00	-.06	-.03	.03	-.06	-.06	.00	-.06	-.03	.03	-.06	-.06	.00
ZGRADE.Fem		.00	.00	.01	-.01	.00	-.01	.00	.00	.00	-.01	.00	-.01	.00	.00	.01	-.01	.00	.00
ImmG1		.15	.13	-.06	.20	.27	.05	.14	.12	-.06	.20	.25	.04	.15	.13	-.06	.21	.26	.04
ImmG1.Fem		.01	.04	-.02	-.02	-.01	.01	.01	.04	-.01	-.01	.01	.01	.01	.04	-.01	-.01	.01	.01
ImmG2		.10	.12	.01	.15	.17	.08	.10	.11	.01	.16	.16	.08	.11	.12	.01	.16	.18	.07
ImmG2.Fem		.01	.00	-.03	-.03	-.02	-.01	.01	.01	-.04	-.03	-.02	.00	.02	.01	-.05	-.03	-.02	.00
Random Effects																			
L3 Cntry		.023	.026	.036	.058	.060	.029	.022	.025	.037	.048	.039	.027	.019	.021	.033	.053	.048	.025
Fem		.008	.003	.007	.006	.006	.006	.008	.003	.007	.005	.005	.006	.008	.003	.007	.006	.006	.007
L2 School		.024	.021	.014	.022	.032	.016	.025	.022	.014	.022	.033	.016	.006	.000	.002	.003	.004	.003
Fem		.006	.000	.003	.004	.004	.003	.006	.000	.002	.004	.004	.002	.740	.698	.773	.852	.785	.903
L1 Student		.744	.702	.780	.853	.788	.905	.746	.704	.777	.857	.789	.905	.019	.021	.033	.053	.048	.025

Table 5C

Generalizability of results to Different Constructs and Alternative Measures of Country-Level Differences in Gender Equality

		L3-Context = GDI						L3-Context =MFRatio University Enrollment						L3-Context =%Muslim					
Fixed Effects		SC	SE	Anx	Util	Int	FP	SC	SE	Anx	Util	Int	FP	SC	SE	Anx	Util	Int	FP
Female (Fem)		-.06	-.14	.09	-.03	.00	-.12	-.06	-.14	.09	-.03	*-.01*	-.13	-.07	-.14	.10	-.03	-.01	-.12
L3-Context		*-.03*	*.00*	*-.03*	*.01*	*-.02*	*.01*	*-.03*	*.00*	*.04*	*-.03*	*.00*	*.00*	.05	.05	.02	.02	.11	.01
L3-Context.Fem	H1	*.01*	*.01*	*.00*	*.00*	*.01*	*-.01*	*.01*	*-.01*	*-.02*	*-.01*	*.00*	*.00*	.02	.00	-.03	.01	.01	.03
L1-SES		.07	.14	-.01	.05	.03	-.01	.06	.13	-.01	.04	.02	-.01	.07	.14	-.01	.04	.03	-.01
SES.Fem		-.02	-.03	*-.01*	-.03	-.03	-.02	-.02	-.03	*.00*	-.02	-.03	-.02	-.02	-.03	-.01	-.03	-.03	-.02
L2-SES		-.01	-.03	-.02	-.05	-.06	*.00*	*-.01*	-.04	-.01	-.05	-.07	*.00*	-.01	-.04	-.02	-.05	-.07	.00
L2-SES.Fem		-.02	-.01	*.02*	-.02	-.02	-.01	-.03	*-.01*	.02	-.03	-.02	*-.01*	-.02	-.01	.02	-.02	-.02	-.01
L3-SES		.08	*.01*	-.12	*.00*	-.10	*.02*	.13	*.02*	-.15	*.07*	*.01*	*.05*	.09	.00	-.14	.01	-.09	.04
`L3-SES.Fem		*-.02*	*-.03*	*.03*	*.00*	*.01*	*-.03*	*-.02*	*-.03*	*.04*	*.01*	*.00*	*-.03*	.00	-.02	.02	.01	.02	-.03
L1-MAch		.81	.67	-.60	.35	.47	.43	.80	.66	-.59	.34	.46	.42	.81	.67	-.60	.35	.47	.43
L1-MAch.Fem		.06	-.04	-.06	.04	.07	.00	.07	-.04	-.06	.04	.08	*.00*	.06	-.04	-.05	.04	.07	.00
L2Mach	H3	-.31	-.07	.19	-.09	-.14	-.09	-.30	-.07	.19	-.08	-.14	-.08	-.30	-.06	.20	-.09	-.14	-.09
L2Mach.Fem		-.04	.03	.04	*-.02*	-.03	*-.02*	-.04	*.02*	.04	*-.03*	-.03	*-.02*	-.04	.02	.04	-.02	-.03	-.02
L3Mach	H4	-.19	-.14	*.01*	-.19	-.19	-.08	-.26	-.16	.02	-.27	-.30	*-.09*	-.16	-.07	.03	-.18	-.11	-.09
L3Mach.Fem		-.06	-.03	.05	*-.03*	-.05	-.01	*-.04*	*-.03*	*.03*	*-.03*	*-.03*	*.00*	-.06	-.03	.04	-.03	-.05	.01
L1-RAch	H5	-.35	-.17	.16	-.13	-.26	-.30	-.35	-.17	.16	-.13	-.26	-.30	-.35	-.17	.17	-.13	-.27	-.31
L1-RAch.Fem		-.02	.02	.06	-.04	-.04	-.02	-.02	.02	.06	-.05	-.04	*-.02*	-.02	.02	.06	-.04	-.04	-.01
L2-RAch	H6	.05	.02	-.03	*.01*	.03	.03	.05	.02	-.03	*.01*	.02	.03	.05	.02	-.03	.01	.03	.03
L2-RAch.Fem		*.01*	*-.01*	-.02	*.01*	.02	.02	*.01*	*-.01*	-.02	*.01*	.02	.02	.01	-.01	-.02	.01	.02	.02
ZGRADE		-.06	-.03	.03	-.06	-.06	*.00*	-.06	-.03	.03	-.06	-.06	*.00*	-.06	-.03	.03	-.06	-.06	.00
ZGRADE.Fem		*.00*	*.00*	*.01*	*-.01*	*.00*	*.00*	*.00*	*.00*	*.00*	*-.01*	*.00*	*-.01*	.00	.00	.01	-.01	.00	-.01
ImmG1		.15	.13	-.06	.20	.27	.04	.14	.12	-.06	.20	.25	.04	.15	.13	-.06	.20	.26	.04
ImmG1.Fem		*.01*	*.04*	*-.01*	*-.01*	*-.01*	*.01*	*.01*	.04	*-.01*	*-.01*	*.01*	*.01*	.01	.03	-.02	-.01	-.01	.01
ImmG2		.10	.12	.01	.16	.17	.08	.10	.11	.01	.16	.16	.08	.09	.11	.01	.15	.16	.07
ImmG2.Fem		*.01*	*.00*	-.03	-.03	*-.03*	*-.01*	*.01*	*.01*	*-.04*	*-.03*	*-.02*	*.00*	.01	.00	-.03	-.03	-.03	.00
Random Effects																			
L3 Cntry		.024	.026	.037	.059	.060	.030	.021	.026	.037	.049	.041	.028	.023	.030	.037	.058	.051	.031
Fem		.008	.003	.007	.006	.006	.006	.008	.003	.007	.006	.006	.006	.008	.003	.006	.005	.006	.006
L2 School		.025	.021	.015	.022	.032	.017	.025	.022	.014	.022	.033	.016	.024	.022	.014	.022	.032	.016
Fem		.007	.000	.003	.003	.004	.003	.006	.000	.002	.004	.004	.002	.006	.000	.002	.003	.003	.003
L1 Student		.746	.702	.785	.855	.792	.904	.746	.704	.777	.857	.789	.905	.745	.704	.782	.851	.788	.906

Note. Fixed and random effects for multilevel analyses (with three levels: L1: individual student; L2 = school; L3 = country) relating gender, gender equality, socioeconomic status (SES), math achievement, and reading achievement to MSC. Included are results based on each of four self-belief/motivation outcomes (Self-Concept , Self-Efficacy, Anxiety, Instrumental motivation, Interest, and Future Plans/intentions) and each of three country-level (L3) measures of gender equality. Variable are described in greater detail in Table 1. The key effects (shaded in grey) are tests of each of the sever a priori hypotheses presented earlier. Values in light blue at not statistically significant (p > .05). All effects are presented as standardized effect sizes (i.e., regression weights in which all first-order effect variables are standardized, M = 0, SD = 1).

Supplemental Material 6. MLwiN Macros

```
NOTE:  Code For Analysis of PISA data in relation to Gender Paradox
effects
Note Module to provide estimates for Paradoxical Gender Gap study
Note: start with a clean model and read in standard dataset
WIPE
LOAD "d:\Dropbox\herb\pisa\pisa2012\MLWIN-MACRO ss-coed L3 Gender Gap 15
outcomes  GGI MANY OUT 8APR2019 v6.ws"

Clear

note: clear old stored models
MWIPE

MARK 0
BATCH 1
weight  1 2  'xnormwt'
Note: Set up basic model with dep variable and level indicators

Note: Add weights
NFMT 1 4
WSET
weight  1 2  c9997
weight  2 2  c9996
weight  3 2  c9995
Note: weights level 1 with standardized wts(1) in "c8"
weight  1 2  'xnormwt'
PREF 0
POST 0
offsets 1
offsets 2
offsets 3
erase  c1091   c1090
mark 1  c1091
mark 1  c1090

Note: Use EXCL to select cases to be used
EXCL 1  c87
NOTE: EXCL 1  c85

RESP  'ZSCMAT'
IDEN 1 c2
IDEN 3 c6
IDEN 2 c7
ESTM 1
EXPA 2
CENT 0
Note: Add variables to be considered
ADDT   'cons'
CENT 0
SETV 3  'cons'
SETV 2  'cons'
SETV 1  'cons'
CENT 0

Note Model 1A Table 2 Effect of GGI, Female & F-GGI Interaction
ADDT 'zGGI'
ADDT  'female'
AddT   'zGGI' 'female'
CENT 0
SETV 3    'Female'
SETV 2    'Female'
SETV 3   'Female'
SETV 2  'Female'
smat 2 0
smat 3 0
```

```
START
Mstor    '\M1a'

Note Model 1B Table 2 Add L1-SES & F-L1SES Interaction
 ADDT  'ZESCS'
 ADDT 'zescs' 'female'
START
Mstor    '\M1b'

Note Model 1c Table 2 Add L2-SES & F-L2SES Interaction
 ADDT  'zL2NZescs'
 ADDT 'zL2NZescs' 'female'
START
Mstor    '\M1c'

Note Model 1d Table 2 Add L3-SES & F-L3SES Interaction
 ADDT  'zL3NZescs'
 ADDT 'zL3NZescs' 'female'
START
Mstor    '\M1d'

Note Model 1e Table 2 Add L1MathAch & F-L1MathAch Interaction
 ADDT  'ZPVxmath'
 ADDT 'ZPVxmath' 'female'
START
Mstor    '\M1e'

Note Model 1f Table 2 Add L2MathAch & F-L2MathAch Interaction
 ADDT  'zL2Mach'
 ADDT 'zL2Mach' 'female'
START
Mstor    '\M1f'

Note Model 1g Table 2 Add L3MathAch & F-L3MathAch Interaction
 ADDT  'zL3Mach'
 ADDT 'zL3Mach' 'female'
START
Mstor    '\M1g'

Note Model 1h Table 2 Add L1ReadAch & F-L1ReadAch Interaction
 ADDT  'ZPVxREAD'
 ADDT 'ZPVxREAD' 'female'
START
Mstor    '\M1h'


Note Model 1i Table 2 Add L2ReadAch & F-L2ReadAch Interaction
 ADDT  'Zl2ZPVxREAD'
 ADDT 'Zl2ZPVxREAD' 'female'
START
Mstor    '\M1i'

Note Model 1j Table 2 Add L1-Year-in-school & F-L1-Year-in-school
Interaction
Addt 'Zgrade'
Addt 'Zgrade'  'female'
START
Mstor    '\M1j'


Note Model 1K Table 2 Add 1ST & 2nd generation immigrant & interactions
with gender-L3MathAch Interaction
Addt 'ImmigG1'
Addt 'ImmigG1'  'female'
Addt 'ImmigG2'
Addt 'ImmigG2'  'female'
START
Mstor    '\M1K'
```

```
Note: Models in Table 3A
Note: this final model is then repeated with different math self-belief
scales
Note: 'ZSCMAT' 'ZMATHEFF' 'ZANXMAT' 'ZINSTMOT' 'ZINTMAT' 'ZMATBEH'
'ZMATINTFC'

Note Model 1K Table 3 with self-efficacy as the outcome
RESP  'ZMATHEFF'
START
Mstor    '\M2b'


Note Model 1K Table 3 with anxiety as the outcome
RESP  'ZANXMAT'
START
Mstor    '\M2c'

Note Model 1K Table 3 with utility/instrumental motivation as the outcome
RESP  'ZINSTMOT'
START
Mstor    '\M2d'


Note Model 1K Table 3 with utility/instrumental motivation as the outcome
RESP  'ZINTMAT'
START
Mstor    '\M2e'

Note Model 1K Table 3 with future plans as the outcome
RESP  'ZMATINTFC'
START
Mstor    '\M2f'


Note: Models in Table 3B
Note: Applying Model K with different Self-beliefs and contextual
variables

Note Model 1K Table 3b with OECD as the contextual variable
RESP  'ZSCMAT'
DELT 'zGGI'
DELT 'zGGI' 'female'
ADDt 'OECD'
ADDt 'OECD' 'female'
START
Mstor    '\M3a'


Note Model 1K Table 3 with self-efficacy as the outcome
RESP  'ZMATHEFF'
START
Mstor    '\M3b'


Note Model 1K Table 3 with anxiety as the outcome
RESP  'ZANXMAT'
START
Mstor    '\M3c'

Note Model 1K Table 3 with utility/instrumental motivation as the outcome
RESP  'ZINSTMOT'
START
Mstor    '\M3d'


Note Model 1K Table 3 with utility/instrumental motivation as the outcome
RESP  'ZINTMAT'
```

```
START
Mstor    '\M3e'

Note Model 1K Table 3 with future plans as the outcome
RESP  'ZMATINTFC'
START
Mstor    '\M3f'



Note Model 1K Table 3c with L3Ach-MFDev as the contextual variable
Calc c830 = ('ZL3devFL4Math' + 'ZL3devFL4Read' + 'ZL3devFL4SCI')/(3 *
.9426)
Name c830 'ZL3devFL4TAch'
RESP  'ZSCMAT'
addT 'ZL3devFL4TAch'
addT 'ZL3devFL4TAch' 'female'
delt 'OECD'
delt 'OECD' 'female'
START
Mstor    '\M4a'


Note Model 1K Table 3 with self-efficacy as the outcome
RESP  'ZMATHEFF'
START
Mstor    '\M4b'


Note Model 1K Table 3 with anxiety as the outcome
RESP  'ZANXMAT'
START
Mstor    '\M4c'

Note Model 1K Table 3 with utility/instrumental motivation as the outcome
RESP  'ZINSTMOT'
START
Mstor    '\M4d'


Note Model 1K Table 3 with utility/instrumental motivation as the outcome
RESP  'ZINTMAT'
START
Mstor    '\M4e'

Note Model 1K Table 3 with future plans as the outcome
RESP  'ZMATINTFC'
START
Mstor    '\M4f'
```