Sex and age differences in 'theory of mind' across 57 countries using the English version of 'Reading the Mind in the Eyes' test

Authors: David M. Greenberg[1,2,3], Varun Warrier[3], Ahmad Abu-Akel[4,5], Carrie Allison[3], Krzysztof Z. Gajos[6], Katharina Reinecke[7], P. Jason Rentfrow[8], Marcin A. Radecki[9], & Simon Baron-Cohen[3]

Affiliations:

[1]Interdisciplinary Department of Social Sciences, Bar-Ilan University, Israel

[2]Department of Music, Bar-Ilan University, Israel

[3]Autism Research Centre, Department of Psychiatry, University of Cambridge, UK.

[4]Institut de Psychologie, Université de Lausanne, Switzerland

[5]School of Psychological Sciences, University of Haifa, Israel

[6]Harvard Paulson School of Engineering and Applied Sciences, Harvard University, USA

[7]Department of Computer Science and Engineering, University of Washington, USA

[8]Department of Psychology, University of Cambridge, UK

[9]Social and Affective Neuroscience Group, IMT School for Advanced Studies Lucca, Italy

Corresponding authors: Email: dmg39@cam.ac.uk (DMG)

Data deposition: Scripts used to analyze the data have been deposited in the Open Science Framework, https://osf.io/hu63x/?view_only=ce5703afb2aa4529bdb0def3beafdc15

Significance Statement

In the seemingly largest study to date on ‘Reading the Mind in the Eyes Test’ (Eyes Test)—a performance task of ‘theory of mind’—we leveraged four unique datasets (total *N* = 312,739), using the English version of the Eyes Test. We found an on-average female advantage across 57 countries. In line with this is a systematic review of translated non-English versions of the Eyes Test identifying an on-average female advantage in eight different languages. Cross-sectional analyses also showed subtle age differences in Eyes Test scores across the lifespan. We conclude there is an on-average female advantage across the majority of countries tested. Future research should investigate this in non-English speakers.

Abstract

The ‘Reading the Mind in the Eyes Test’ (Eyes Test) is a widely used assessment of ‘theory of mind’. The NIMH Research Domain Criteria (RDoC) lists it as one of two recommended tests for individual variation in ‘Understanding Mental States’. Previous studies have demonstrated an on-average female advantage on the Eyes Test. However, it is unknown if this female advantage exists across the lifespan and across a large number of countries. Thus, we tested sex and age differences using the English version of the Eyes Test in adolescents and adults across 57 countries. We also tested for associations with sociodemographic and cognitive/personality factors. We leveraged one discovery dataset (*N* = 305,726) and three validation datasets (*N*s = 642; 5,284; and 1,087). The results show that: (i) there is a replicable on-average female advantage in performance on the Eyes Test; (ii) performance increases through adolescence and has a shallow decline across adulthood; (iii) the on-average female advantage is evident across the lifespan; (iv) there is an on-average female advantage in 36 out of 57 (63%) countries; (v) a systematic review shows a female-advantage on translated (non-English) versions of the Eyes Test in 12 of 16 countries; (v) empathy-systemizing 'brain types' predicted Eyes Test above and beyond sex differences; and (vii) exploratory analyses at the country-level showed that the female advantage is negatively linked to ‘Prosperity’ and ‘Autonomy’, and positively linked to ‘Collectivism’. We conclude that the on-average female advantage on the Eyes Test is observed across ages and most countries.

Introduction

‘Theory of mind’ (ToM) is the ability to attribute mental states to oneself and others, in order to make sense of human behaviour and to predict it (1). Since the 1980s, ToM has become central to the study of human development, particularly the development of social perception and social cognition, and to understanding clinical conditions such as autism, conduct disorder, personality disorders, anorexia, and schizophrenia (2–8). ToM is also a central focus of research in comparative psychology, addressing the question of whether ToM is unique to humans (9), to research in neuropsychology, addressing how brain lesions affect ToM (10–12), and to social neuroscience, testing the biological and social factors that influence ToM (11, 12).

There is evidence that ToM follows consistent developmental patterns during childhood, with a progression through different stages. Although the precursors of ToM in infancy are debated (13), some suggest that precursors of ToM are evident between 9 to 15 months of age in joint attention behaviours such as gaze-following, showing behaviours, and gestures such as pointing to share interest (‘protodeclarative pointing’) (14–17). It is notable that autistic children at the earliest point they can be diagnosed show delays or deficits in both joint attention and pretend play, and in later developmental milestones of ToM (18).

In the second year of life, young typical children understand the mental states of goals and desires of others (9), and at about 4 years old, children understand that another person can have a different, false belief, (so-called first-order ToM) (9, 10). By around 5 to 6 years of age, children understand what someone is thinking about another person’s mental state (second-order ToM) (20). Later, children also recognize *faux pas*, which is evident around 9–11 years of age and refers to the ability to understand and recognize situations in which someone has said something inappropriate that a listener either did not need to know or which could be hurtful (21). This is relevant to ToM because it is a clear sign that children are monitoring what others know or need to know, and that they have feelings that could be hurt. Autistic children are delayed in passing *faux-pas* tests and autistic adults report finding it hard to just what is socially appropriate to say or picking up on when or why someone might have taken offense and what was said (22). ToM abilities continue to develop well in late adolescence (23). Whether developmental progression is identical across countries is debated, with some suggesting its development occurs uniformly across cultures (21), while others suggest it is culture-specific (16).

Multiple performance tasks have been developed to measure first-order ToM, including the Emotional Triangles (26) and as reviewed above, False Belief tasks (27). One of the most widely used tasks in the past two decades, particularly in the study of adults' ToM, is the 'Reading the Mind in the Eyes Test' (Eyes Test) (28). The Eyes Test is a paper-and-pencil or online performance task where respondents are presented with 36 pictures of the eye region of a human face and asked to indicate which of four word choices best describes what the person in the picture is thinking or feeling. Reduced performance on the Eyes Test has been reported in autistic individuals (29), those with eating disorders (4), personality disorders (7), schizophrenia (5), substance abuse disorders (30), or dementia and Alzheimer's disease (31). Patients with known brain lesions in the amygdala, and inferior frontal gyrus show acquired deficits on the Eyes Test (10, 11). Autistic people and their siblings both show reduced brain activity in these regions during Eyes Test performance in an fMRI scanner (32, 33). These clinical differences suggest that the Eyes Test may be one measure with which to investigate differences in social processes both in individuals with neurodevelopmental and psychiatric conditions and in the general population. Accordingly, The NIMH Research Domain Criteria (RDoC) lists the Eyes Test as one of two recommended tests for the measurement of individual differences in 'Understanding Mental States' (https://www.nimh.nih.gov/about/advisory-boards-and-groups/namhc/reports/behavioral-assessment-methods-for-rdoc-constructs.shtml).

There is evidence that both biological and social factors contribute to individual differences in performance on the Eyes Test. In terms of biological factors, performance on the Eyes Test is partly genetic, with a twin heritability of 28% (95% CIs: 13% – 42%), and a SNP-based heritability of 5.8% (95% CIs: 4.5% – 7.2%) (34). Performance on the Eyes Test is also associated with prenatal testosterone (35), current testosterone (van Honk et al, 2011),, and with intranasal oxytocin administration (36), implicating biological mechanisms that influence performance on ToM tasks (36–39). In terms of social variables, in adolescents, individual differences in performance on the Eyes Test are associated with smartphone usage, texting, and engaging in fantasy play (41).

Convergence across studies and meta-analyses show robust sex differences on the Eyes Test, with an on-average female advantage (24, 31, 42, 43). The female advantage could be due to the same set of biological factors that contribute to individual differences in, for example, prenatal testosterone (which is on average higher in males than females) (35), or due to partly different genetic architecture between males and females (34). In terms of social factors, one potential explanation for the on-average female advantage on the Eyes Test (at least in adolescents and adults) is the gender-intensification theory, where the female advantage is seen as partly due to expected gender roles (44). Also relevant is Wood and Eagly’s (45) conceptualization of gender as a biosocial construct that results from complex interactions between biology and experience. It is important to note that an on-average female advantage is not necessarily found across all ToM tasks (44), and some argue that the Eyes Test does not capture ToM but rather emotion recognition (47). Emotion recognition is an important part of ToM, and the Eyes Test captures aspects beyond emotion recognition, as some of the mental states tested include items that are epistemic mental states (such as planning or scheming). When ToM is considered inclusive of emotion recognition, the evidence for the female advantage extends well beyond the Eyes Test, , with study samples ranging from 3,000 to 100,000 participants across the lifespan (48–52).

At the geographical level, there are sex differences in personality traits and preferences to study or work in STEM (science, technology, engineering, and mathematics), even in countries that have lower gender inequality (49–51). This so-called gender-equality paradox (57–60) suggests that any residual gender differences in societies with less gender discrimination (e.g., that have moved closer towards gender equality) may reflect partly biological factors. These, in turn, might reflect individual differences that are highly specific (such as greater attention to the eye region of the face, with there being a female advantage in facial recognition; (61) or that are much broader (such as a stronger interest in people than in objects, with there being a greater female social interest (60). Studying sex is important given that conditions such as autism and schizophrenia, where scores on the Eyes Test are different from the general population, also have a marked sex bias (49, 50). A more comprehensive investigation of the correlates of sex differences on ToM tasks will enable us to better understand the sex differences in conditions such as autism and schizophrenia. There is thus a need for a large-scale, robust study to test these variables definitively.

In contrast to the sex effects on the Eyes Test, age effects on the Eyes Test are less clear. Some studies suggest that scores fluctuate during adolescence, but are stable across adulthood (64). Other studies are contradictory, showing a decrease in scores on the Eyes Test across adulthood (65), which has been replicated with other ToM tasks (66), and others showing an increase in scores across adulthood (67). To our knowledge, there has been no comprehensive investigation of normative age trajectories in performance on the Eyes Test. For instance, it is unknown if there are large age-related declines in performance on the Eyes Test, and if there are sex differences in this decline. Although the Eyes Test is widely used, there are gaps in our knowledge, mainly in whether the on-average female advantage generalizes across all countries, and whether there are robust age trends. These gaps in the literature are largely because of a historical reliance on small sample sizes and relatively homogenous samples in terms of both geography and age.

We address these gaps in the literature by using a large and geographically diverse sample to test sex and age differences on the English version of the Eyes Test (the discovery dataset). In addition, we leverage three separate samples (validation datasets A, B, and C) to replicate and extend results from the discovery dataset (validation datasets A and C used the full 36-item version of the Eyes Test and validation dataset B used an 18-item version) (**Methods**). In each of the four studies, participants were asked to indicate their sex, not their gender, although we acknowledge that in common parlance the terms sex and gender are often used interchangeably. We test for: (i) on-average sex differences; (ii) on-average age differences; (iii) associations with demographic variables (including educational attainment), personality/cognitive variables, including the Big Five personality traits (68), and empathizing-systemizing (E-S) cognitive profiles (also referred to as E-S 'brain types') (56). These latter profiles equate to D-scores, the standardized difference between a person's score on the Empathy Quotient (EQ) (22) and Systemizing Quotient-Revised (SQ-R) (70). E-S brain types have been shown to have a brain basis (71–77). Using country-level data, we further test for: (iv) on-average sex-differences across countries, and, as exploratory analyses, (v) the association between country-level sociodemographic factors via Political, Economic, Social, and Health (PESH) indicators (78, 79), including the association between gender equality and sex differences on the 'Eyes Test'.

The analyses using PESH indicators are exploratory, and whilst they represent an avenue to understand the possible social mechanisms of the associations, this is not the main aim of our study. Our aim is simply to identify whether sex differences are observed across countries. We distinguish country from culture and do not use the terms synonymously here. For instance, India has multiple cultures but is considered a single country/nation. We also conduct a systematic review of studies that used translated versions of the Eyes Test to determine whether sex differences that emerge from our datasets also emerge in non-English speakers and non-English versions of the Eyes Test. **Fig. 1** provides a schematic diagram of the study and sample characteristics are presented in **Table S1**.

Results

Is there an on-average sex difference on the Eyes Test?

For our main analysis, we conducted Bayesian multi-level analysis, using a normal prior: N (0,1) on the discovery dataset. We estimated sex differences and age, and the prior was applied to Eyes Test scores. Posterior estimates identified an on-average female advantage after including age as a covariate and country as a random intercept (β = .17; SE = .00; 95% CIss = [0.16, 0.18]). For each of the three validation datasets, we conducted the same analysis, however, without country as a random intercept since there was not sufficient country-level data in each of the validation samples. Beta estimates equal to 0 provide an estimate of the plausibility that there is no sex difference; beta estimates above 0 indicate a female advantage and beta estimates below 0 indicate a male advantage. As seen in Table 1, beta estimates are all above 0 and range from .17 (SE = .00, 95% CIss = [.16, .18]) to .27 (SE = .00, 95% CIs = [.22, .32]). As seen in Fig 2, conditional effects show no overlaps in the 95% credible intervals exceeding the 17% overlap threshold for evidential support (80). Taken together, the results shown in **Table 1** and **Fig. 2** provide robust evidence that females outperformed males across all four datasets. Reliability on the Eyes Test in each of the four datasets showed acceptable-to-good reliability using McDonald's Omega total (ω_t) and Omega hierarchical (ω_h) (**Table 1**).

With the Bayesian model of the discovery dataset, we examined sex differences within each of the 57 countries that met our inclusion criteria (**Methods**). The sample size per country ranged from *n* = 112 (Vietnam) to *n* = 176,402 (United States) (sample sizes for each country are presented in **Table S2**). There was only one country in which females did not have a higher descriptive mean score on the Eyes Test than males (i.e., Colombia) (**Table S2**). As before, beta estimates above 0 indicate a female advantage, while beta estimates below 0 indicate a male advantage. As can be seen in **Fig. 3** and **Table S2**. Analysis of credible intervals showed that there was a female advantage in 36 of 57 countries (63%). That is, 36 countries had a lower bound credible interval ≥ 0 providing evidence for a female advantage, while no countries have a higher bound credible interval ≤ 0, which indicated that no countries had evidence for a male advantage. Facet plots with conditional effects for each sex for each country are presented in **Fig S1**.

Next, we examined reliability of the Eyes Test within each of the 57 countries in the discovery dataset. As can be seen in **Fig. 4** and **Table S2**, there was acceptable-to-good reliability for 56 of the 57 countries, with (ω_t) ranging from 0.68 to 0.92. However, one country, Colombia, had poor reliability with (ω_t) = .49. Since there was a larger sample of countries (57) than datasets (4), we decided to correlate the beta of each country with the reliability estimates. A high correlation would indicate an effect of reliability on the betas. Column-vector correlations from **Table S2** between the beta and reliability columns was .27, and when first conducting Fisher's *r* to *z* transformation prior to performing the correlation, it was .18, suggesting minimal effect on the betas.

Our results related to sex differences from the discovery and validations datasets were limited, because we were reliant on proficient speakers of English taking the English version of the Eyes Test. To address this limitation in the data, we conducted a Systematic Review to identify on-average sex differences in non-English versions of the Eyes Test. There were 16 studies included in the review, with 10 different translations of the Eyes Test. Our study-selection process is based on the PRISMA model (**Methods**) and is presented in **Fig. S1**. Out of 16 studies selected to review, 12 studies (that include eight translated versions of the Eyes Test) showed a significant female advantage, and the remaining four studies showed a descriptive female advantage that did not reach statistical significance. The study characteristics are presented in **Table S3.** This shows that an on-average female advantage tends to be found in translated versions of the Eyes Test and validates our findings, which were based on the English version of the Eyes Test.

Are there on-average age differences on the Eyes Test?

The large sample in the discovery dataset enabled us to check for on-average sex differences within each age year (i.e., 16, 17, 18, 19 … 70). Our results show that there is indeed a persistent on-average female advantage in each age year (i.e., 16, 17, 18, 19 … 70) (**Table S4**). In terms of age itself, results from the Bayesian multi-level model demonstrated that age had a minimal effect within the linear model (*β* = .03; SE = .00; 95% CIss = [0.02, 0.04]) (means, *SD*s, and beta estimates for each age from 16 to 70 are presented in **Table S4**). To identify peaks (i.e., inflection points) in the age trends, we performed a constrained non-linear regression analysis separately for females and males within a frequentist model, with a trimmed estimator to obtain a more robust statistic. In terms of age trends for each sex, for females, there were thresholds at 20.25 years of age (*SE* = .43; 95% CIs = [19.20, 20.69]) and at 49.82 years of age (*SE* = 7.96; 95% CIss = [41.18, 63.18]). There was evidence for an increase in performance on the Eyes Test from age 16 to 20.25 years old (*β* = .40, *SE* = .07, 95% CIss = [.26, .53]), a shallow decline from age 20.25 to 49.82 years old (*β* = -.013, *SE* = .002, 95% CIs = [-.014, -.008]), and then a further decline—by a factor of 2—from age 49.82 (*β* = -.028, *SE* = .047, 95% CIs = [-.159, -.018]).

For males, there were thresholds at 20.48 years of age (*SE* = .65; 95% CIs = [19.79, 21.58]) and 58.14 years of age (*SE* = 3.36; 95% CIs = [56.89, 67.06]), with evidence for an increase from age 16 to age 20.48 (*β* = .43, *SE* = .08, 95% CIs = [.25, .52]), a shallow decline from age 20.48 to age 58.14 (*β* = -.009, *SE* = .002, 95% CIs = [-.012, -.006]), and then a further steeper decline—by a factor of 8—from age 58.14 (β = -.069, *SE* = .134, 95% CIs = [-.486, -.054]) (**Fig. 5**). In **Fig S3**, we provide facet plots for each country, showing age trends from 16 to 70 on the Eyes Test, for females and males separately. These facet plots are based on LOESS regression for each country by sex. Overall, the plots show a similar trend (i.e., Eyes Test scores decline throughout adulthood). Those countries that do not show this trend are countries that have smaller relative sample sizes (e.g., Nigeria).

Are there sociodemographic and cognitive/personality associations with scores on the Eyes Test?

We expanded our main Bayesian multi-level model by adding sociodemographic and cognitive/personality variables available in each dataset as covariates (**Methods**). First, sex retained an effect on the model in each dataset, even after adding the covariates (*βs* = .09 to .20) (**Table S5**). In terms of sociodemographic variables, only education had a positive effect on performance in the discovery (β = .11, SE = .03, 95% CIs = [.04, .18]) but not in the validation datasets. There was no evidence that income was a predictor across the datasets. Web usage had a positive effect on the model of the discovery dataset (*βs* = .15, SE = .03, 95% CIs = [.08, .22]). In terms of cognitive/personality variables, D-scores (the standardized difference between scores on the Empathy Quotient (EQ) and Systemizing Quotient (SQ-R)), and the basis of 'cognitive-profile' calculations (high scores indicate a bias towards systemizing and low scores indicate a bias towards empathizing) (**Methods**) had a negative impact on performance in each of the three datasets in which it was included: (*βs* = -.13 to -1.06). The Big Five personality traits did not have an effect on performance. Overall, the effect of D-scores had the greatest effect on the performance and sex had the second-greatest effect, which underlines the importance of understanding the role of D-scores on performance of the Eyes Test in future research.

Are on-average sex differences on the Eyes Test associated with country-level factors?

Finally, we tested country-level correlates that may shed light on the geographical differences in the magnitude of the female advantage across countries. Since each participant confirmed they understood each word-descriptor of each item of the Eyes Test, we conducted an exploratory ecological-correlation analysis at the country level to find associations between on-average sex differences on the Eyes Test and country-level variables. Specifically, we leveraged 16 different country-level variables that outline PESH indicators (79, 81), including gender-equality indices (**Methods**). Initial analysis showed that Pakistan was an outlier for the on-average sex differences and Global Gender Gap Index (GGGI) (Mahalanobis distance = 9.00), so we removed it from further analyses.

The 16 PESH indicators were available for 52 of the 56 remaining countries. To reduce the number of indicators, we performed a principal-component analysis with varimax rotation. The KMO test for sampling adequacy was .72, and there was a clear elbow on the scree plot at four components, suggesting that we retain three components, which together accounted for 86% of the variance. The component loadings are presented in **Table S6**. Component 1 outlined indicators related to 'Prosperity' (e.g., income index, Human Development Index, global creative index), component 2 outlines 'Autonomy' (e.g., intellectual and affective autonomy, democracy index, and egalitarianism), and component 3 outlines 'Collectivism' (e.g., harmony, and a negative loading for mastery).

We then conducted a Bayesian multi-level model where each of the three components were regressed onto the country-level beta estimate on the Eyes Test. To account for the fact that the different countries have different sample sizes, we added each country's sample size as weights. All three components demonstrated an effect on the model. Both the Prosperity (*β* = -.15, SE = 0.00, 95% CIs = [-.15, -.14]) and the Autonomy (*β* = -.11, SE = 0.00, 95% CIs = [-.12, -.11]) components were negatively associated with beta estimates (**Table S7**). The Collectivism component was positively associated with beta estimates (*β* = .08, SE = 0.00, 95% CIs = [.08, .09]). This suggests that the more prosperous and autonomous a country is, the smaller the female advantage, and the more collectivist a country is, the greater the female advantage.

Discussion

We confirm an on-average female advantage on the widely used 'Reading the Mind in the Eyes' Test (Eyes Test) across four samples assessed with the English version of the test. We show that this on-average female advantage persists across the lifespan from 16 to 70 years of age. Our results also show that the on-average female advantage on the English version of the Eyes Test was evident in 36 of the 57 countries that we observed (in those individuals whose primary or second language was English). There was no country where males scored significantly higher on the Eyes Test than females. Our Systematic Review of translated versions of the Eyes Test also shows a female advantage that reached statistical significance across 12 of 16 studies.

Effect sizes can be meaningless without context (82), particularly with complex phenomena like sex differences (83). The effect of the sex difference across the discovery and validation samples could be interpreted as a small or very small effect (84). However, recent theory and research has adopted new benchmarks for interpreting effect sizes which suggest that even small effects can be consequential in the long run (85). Indeed, there is a growing consensus to accept small effects in psychology as the norm and to acknowledge that small effect sizes can have substantial consequences for human behaviour. The cumulative evidence from our data that the sex effect appears in multiple countries strengthens the importance for its study. Further research is needed to evaluate how this on-average sex difference maps onto real-life outcomes.

We cannot determine causation from our data, as our study does not investigate mechanisms. The robust on-average female advantage on the Eyes Test across countries may have both biological and social determinants. For instance, the gender-intensification theory (44) suggests that the observed sex differences might be partly explained by expected gender roles to which children, adolescents, and adults have increased pressure to conform. In terms of biology, there is a significant negative correlation between prenatal testosterone and scores on the Eyes Test (35). Additionally, whilst SNP heritability for the Eyes Test is identical between the sexes, the genetic correlation between the sexes in adulthood is modest and statistically less than 1 (with 1 being the maximum genetic correlation) (43). This suggests that there may be different genetic pathways underlying the development of ToM between the sexes. In terms of biosocial explanations, it is possible that early sex differences stem from biological factors, but are maintained or amplified by social factors that are prevalent in the countries we observed (86). These potential mechanisms should be investigated in larger studies to better understand what generates these sex differences.

Our findings on age-related trajectories add to prior evidence suggesting age differences in ToM throughout the lifespan (87, 88). For females, our results showed peaks in Eyes Test scores at 20 years of age, with an additional inflection point at 50. For males, there was also a peak at 20 years of age, but an inflection point at 58. The decline in both females and males during later adulthood replicates and extends a previous meta-analysis (*N*s = 790 younger adults and 672 older adults), which showed poorer performance on multiple ToM tasks by older adults compared to younger adults (65). The differences across the lifespan for females and males, coupled with the sex differences on the Eyes Test raises questions for future research on the role of hormones and their contribution to the development of ToM during adolescents and shallow decline in adulthood.

Our findings also showed that sociodemographic and cognitive/personality factors play a role in performance on the Eyes Test. In particular, the effect of D-scores predicted Eyes Test scores above and beyond the effect of sex.. Similarly, a study of more than 650,000 people found that D-scores accounted for 19 times more of the variance in autistic traits than that of sex and other demographic variables (69). Therefore, evidence is accumulating to show that D-scores play a more important role in different aspects of human cognition than does sex. Separately, the effect for web usage in predicting Eyes Test scores was larger than the effect of sex. Whilst this points to sociocultural factors that influence Eyes Test scores, a likely explanation is that it is simply a reflection of fluency with the computer as a tool.

In our exploratory country-level analyses, we observed that the magnitude of the on-average female advantage in each country had associations with country-level PESH indicators. Specifically, the female advantage on the Eyes Test was positively correlated with cultural values rooted in the collective and negatively correlated with prosperity and autonomy. Prior research shows that individuals with lower social status demonstrate more care and concern for others' thoughts and feelings (89, 90), attend more closely to the social cues of their partner (91), and are more likely to give financially to others, including giving a higher proportion of their income to charity (92). However, it is unclear how the association between social status and concern for others (93) translates to the on-average sex difference on the Eyes Test. A replication and extension is needed. This points to the added value that country-level analyses bring to our understanding of the Eyes Test and ToM more generally. Taken together, the on-average female advantage appears to reduce across more progressive and Westernized countries. These findings lay an initial basis from which future work can build. Rigorous cross-cultural research is needed to shed light on these questions.

Our results provide robust validation for the psychometric properties of the Eyes Test. This is supported by acceptable-to-good reliability of the Eyes Test across datasets and countries. Since the Eyes Test is widely used and is listed as a recommended test for measuring individual differences in 'Understanding Mental States' by the NIMH RDoC, establishing such a validation is useful for researchers intending to use this test. One concern that has been raised about the Eyes Test is that all 36 stimuli are of White faces. Research findings addressing this concern are mixed. One study found no cultural differences in performance on the Eyes Test when using stimuli of White faces and stimuli of Black faces, regardless of the race of the participant (94). However, some research shows that face processing, more generally, may be biased towards own-race faces (95). Therefore, although studies of the Eyes Test when translated to different languages (96–100) demonstrate that the Eyes Test is a suitable measure for the study of social processes in different geographic contexts, more cross-cultural studies are needed.

Our study had several limitations. First, our study only included English speakers and the English version of the Eyes Test, which limits our conclusions across countries. These English speakers all had access to a computer suggesting that samples from some countries may be biased and not representative of the demographics of the population. Hence, it is unclear to what extent the sex differences identified in the samples in our studies are truly representative of the average sex difference in some countries, thereby limiting the interpretation of the cross-country generalizability and of the PESH analyses. Second, whilst we have endeavoured to examine translated versions of the Eyes Test in our systematic review, our study may still be limited in its geographical reach. This may be particularly problematic for countries where English is not spoken widely or with more modest internet and computer access. Additionally, we have not explicitly tested if performance on the Eyes Test varies between cultures. Future research should explore if the on-average female-advantage is replicated in more traditional societies with minimal exposure to Western culture (101). Third, analysis of age trends was based on cross-sectional rather than longitudinal data, and the age range for the adolescent sample (i.e., ages 16 to 19) was much narrower than for older ages (e.g., 40 to 49). Fourth, ToM was assessed by only a single measure (albeit a widely used and reliable performance measure). Further research is needed to investigate if other ToM tasks demonstrate a female advantage across ages and countries. Fifth, the adult Eyes Test has not been extensively validated on datasets including participants aged less than 16 years old, and therefore we were unable to test for individual differences in scores in early to middle childhood and early adolescence. Sixth, causation about the observed sex differences in our study cannot be inferred. All these limitations should be addressed in future research.

Although three of the four datasets we used asked participants ‘what is your sex’ rather than ‘what is your gender’, several of the datasets included answer choices that were non-binary, including “non-binary’, ‘transgender’, and ‘other’. Furthermore, the question on sex did not specify biological sex assigned at birth. This may have caused confusion among transgender participants and participants who identify as non-binary. Because of the lack of clarity in the question that was administered, we decided not to include transgender participants in our analysis to test how transgender individuals score on average on the Eyes Test. Furthermore, we do not make any assumptions about how transgender or non-binary individuals may have responded, and therefore have not speculated about the possible effects of these sampling choices on the results. Our recommendation for future studies is to ask at least two distinct questions to participants: ‘What is your biological sex assigned at birth?’ (with answer choices ‘female’, “male’, and ‘intersex’) and ‘What is your gender identity?’ (with answer choices that may include ‘cisgender’, ‘transgender’, ‘agender’, ‘gender-non-conforming’/’non-binary’/’genderfluid’, ‘genderqueer’, and ‘other’).

In conclusion, in one of the largest studies to date on ToM, we found robust evidence in support of an on-average female advantage in ToM using the widely used Eyes Test, and were able to replicate the findings in three additional and diverse datasets. The on-average female advantage was present in every age year across the lifespan. We look forward to further research exploring the biological and social determinants of this effect, and how these interact.

Methods

Ethics statement

Ethical approval for the full study protocol of the discovery dataset was provided by the IRB at Harvard University. Ethical approval for the full study protocol of validation dataset A was provided by the Psychology Research Ethics Committee at the University of Cambridge. Validation datasets B and C were given ethical approval to be used as secondary data by the review board at Ethical and Independent Review Services (http://www.eandireview.com). All participants in each dataset provided informed consent.

Discovery dataset

Participants and procedures

More than 460,000 volunteer participants, who were English-speaking, completed an English version of the Eyes Test and demographic questions at www.labinthewild.org from February 2013 to May 2019 (102). Participants were asked “If you are not a native speaker of English, did you recognize all the words used to describe emotions?” with four answer choices: 1 = ‘I am a native speaker of English’; 2 = ‘I am not a native speaker, but I recognized all the words used to describe emotions in the study’; 3 = ‘I recognized almost all the words used to describe emotions in the study’; 4 = ‘I recognized only some of the words used to describe emotions in the study. For the purposes of this study, we only included participants who indicated either 1 or 2 on the native-English-speaking question. After completing the question items, all participants received feedback about their scores on the Eyes Test.

The Eyes Test has not been fully validated in age groups under 16 years old, as there has only been limited research on the child version of the Eyes Test (103–105), therefore we did not include participants under this age. We did not include people aged above 70 years old, because we lost statistical power above this age group. Including a large age range from 16 to 70 years old enabled us to capture ToM development that is suggested to occur in late adolescence (23). Since the question items across the discovery and validation datasets specified "sex" and not "gender", we did not include individuals who identified as non-binary. This left 305,726 individuals for analysis aged 16 to 70 (*M* = 29.57, *SD* = 11.80). 142,696 (48%) were female and the majority of participants were from the United States (*n* = 180,293; 62%) and 30,898 from the United Kingdom (11%). However, because of the large sample size, there was a substantial number of participants from other countries allowing for cross-cultural analysis (**Table 2**).

Measures

Participants first completed demographic items and then completed the 36-item Eyes Test (28) (**Fig. 6**). Demographic items included sex ("What is your gender" with answer choices: 1 = female; 2 = male; and 3 = non-binary [2019 and after]/it's complicated [prior to 2019]"); age (0 to 123), education ("What is the highest level of education you have received or are pursuing?" with answer choices: 1 = pre-high school; 2 = high school; 3 = college; 4 = masters; 5 = PhD); web usage ("How often do you use a computer?" with answer choices: 1 = Once a week or less; 2 = A few times a week; 3 = A couple of hours most days; 4 = Many hours on most days); country living in now ("In what country have you spent most of the past five years?") and country lived in during childhood ("In what country did you live most of your childhood? (Please pick one that influenced you the most if you grew up in more than one")); and self-reported ability to recognize emotions of others (“Compared to your family and friends, how good are you at reading people’s emotions?”) with answer choices: 1 = much worse; 2 = slightly worse; 3 = about the same; 4 = slightly better; and 5 = much better.

Validation dataset A

Participants and procedures

Between April 2007 and January 2017, 642 participants completed the Eyes Test at www.cambridgepsychology.com. Of those who indicated, 422 (66%) were female, and the sample ranged in age from 18 to 70 years old (*M* = 37.07, *SD* = 12.51). Participants were predominantly from Europe, including 333 (52%) from the United Kingdom. There were no questions about English comprehension of the English language in any of the validation datasets and no feedback about scores given to participants. Results on sex differences from a smaller sample (*N* = 320) in this dataset was previously published, and its focus was on sex differences on the Eyes Test in autism (29).

Measures

Each measure presented to participants was in English. Participants are asked “What is your sex” with two answer choices: female and male. All participants completed the Eyes Test, and 639 participants also completed the 40-item Empathy Quotient (EQ) (22), the 75-item Systemizing Quotient-Revised (SQ-R), and the 50-item Autism Spectrum Quotient (AQ) (106). The EQ and SQ-R (107) allow for the calculation of D-scores, which is the standardized difference between EQ and SQ-R scores. We followed the procedure established previously for calculating D-scores (70). To calculate the D-score for each participant, we first standardized the EQ and SQ-

R scores across the whole sample (including both males and females) based on means from the typical population in the sample: S = [(SQ-R−<SQ-R>)/150 and E = (EQ−<EQ>)/80]. That is, we first subtracted the typical population mean (denoted by <. . .>) from each individual's scores, and then divided this by the maximum possible score (150 for the SQ-R, and 80 for the EQ). The D-score is defined as follows: D = S – E. D-scores are often used to provide classifications of five categorical cognitive profiles (sometime referred to in the literature as "brain types"), but since we only used D-score in linear regressions, we did not specify cognitive-profile classifications.

Validation dataset B

Participants and procedures

Between March and November of 2016, 5,284 participants completed an abbreviated version of the Eyes Test at www.musicaluniverse.org. Participants completed a battery of measures for a larger study on music and personality that involved listening to audio excerpts. The sample was geographically diverse with most of the participants from the United States (n = 1,871, (36%) and the United Kingdom (n = 793, 15%). Of those who indicated, 2,947 (56%) were female and age ranged from 16 to 70 years old (M = 33.73, SD = 11.76).

Measures

As part of the battery, participants completed an 18-item version of the Eyes Test, which included the first half of stimuli of the full Eyes Test. The rationale for administering a brief 18-item version, rather than the full 36-item version, was to prevent participant fatigue. This version is strongly correlated with the full version of the test (r = .84, p < .001, N = 642 from validation dataset A). Participants were asked "what is your sex" with four answer choices: female, male, transgender, and other. We only included participants in the analysis who selected female or male. Participants completed a brief measure of the Big Five personality traits, the Ten-Item Personality Inventory (TIPI) (108), and the 5-item Satisfaction with Life Scale (SWLS) (109). A subsample of participants also completed the 40-item EQ and the 25-item short version of the SQ (110). D-scores were calculated using the same procedure as in validation dataset A.

Validation dataset C

Participants, procedures, and measures

Participants in this dataset were users of the same data-collection platform as validation dataset B. However, the only difference was that instead of completing the abbreviated version of the Eyes Test, participants in this sample completed the full version of the Eyes Test. All remaining measures were the same as in validation dataset B. There was a total of 1,087 participants. Of those who indicated, 393 (58%) were female, and the sample ranged in age from 16 to 70 years old (*M* = 33.98, *SD* = 11.97). Sample characteristics for all four datasets are presented in **Table S1**.

Statistical analyses

In the initial analysis, to investigate that the country-wise results are not affected by differing reliabilities of the Eyes Test, we calculated both total McDonald's Omega (ω_t, a measure of the total reliability of both the general and the group factors) and hierarchical McDonald's Omega (ω_h, a measure of reliability of only the general factor) in all the samples and by country in the discovery dataset.

In the main analysis of sex differences, to accommodate both the individual-level and country-level data in the discovery dataset, we adopted a Bayesian multi-level model that fit all the data, and then showed the posterior distribution of the estimated effect sizes of the sex difference, along with the other parameters being estimated. In the Bayesian multi-level analyses, we specified sex as a fixed variate, age as a covariate, and countries as random intercepts and used a normal prior of (0,1). This was conducted using the *brms* package in R version 4.1.2. For the validation datasets, we also conducted Bayesian multi-level analyses using a normal prior of (0,1), with no random intercepts since the validation datasets did not have enough country-level data.

For analysis of age differences in the discovery dataset, we relied on results from the Bayesian multi-level model above. We also performed constrained non-linear regression analysis using a frequentist approach (from 16 to 70 years) to identify peaks and inflection points of age trends.

To test for cognitive/personality and sociodemographic variables that are associated with scores on the Eyes Test, we added to the main Bayesian multi-level models in each of the datasets by adding variables that were available in each dataset as covariates: sex, age, education, different country of birth and web usage in the discovery dataset; sex age, AQ, and D-scores in the validation A dataset; sex, age, education, income, D-scores, openness, conscientiousness, extraversion, agreeableness, neuroticism, and life satisfaction in the validation B dataset; and sex, education, income, D-scores, openness, conscientiousness, extraversion, agreeableness, neuroticism, and life satisfaction in the validation C dataset.

The large and geographically diverse nature of the discovery dataset gave us the opportunity to test sex differences on the Eyes Test across countries. We determined the participant's country location based on the country they indicated they are living in now ("In what country have you spent most of the past five years?"). Since there are no standards for power analysis in Bayesian modelling to determine the number of countries to retain, we relied on a power analysis using G*Power which suggested a total sample size of $N = 107$ to test for two predictors with an effect size of .15 and 95% power. Fifty seven countries met this criterion.

We tested country-level correlations with sex differences using the Political, Economic, Social, and Health (PESH) framework, which has been previously established and tested successfully in research on geographical psychology (79, 81). For political indicators we used the EIU Democracy Index from 2018 (https://www.eiu.com/topic/democracy-index) and the IEP Global Peace Index (GPI) from 2019 (http://visionofhumanity.org/app/uploads/2019/06/GPI-2019-web003.pdf). For economic indicators, we used the Income and Education subindices of the United Nations (UN) Human Development Index from 2017 (http://hdr.undp.org/en/data). For social indicatorsm we used the Global Creativity Index (GCI) from the Martin Prosperity Institute report in 2015 (http://martinprosperity.org/tag/creativity-index/), the Global Gender Gap Index (GGGI) of the World Economic Forum from 2017 (https://www.weforum.org/reports/the-global-gender-gap-report-2018), the Gender Development Index (GDI) of the UN from 2017 (https://hdr.undp.org/gender-development-index#/indicies/GDI), and Schwartz's seven culture/society-level value orientations, which are suitable for comparing countries (111) and which were derived from his prior theory on human values (112). Finally, for health indicators, we used the Life Expectancy subindex of the Human Development Index from 2017. We performed a principal-component analysis (PCA) with varimax rotation on the country-level variables and performed a Bayesian multi-level model while specified the resultant components of the PCA and regressed them onto the country-level beta estimates from sex differences, while adding the sample size of each country as weights.

To test if our country-level samples are representative of their countries, we correlated country-level web usage as indicated from self-report in the discovery dataset with the percentage of people in each country who have access to the internet (gained from Statista and Global Digital Insights) which showed no correlation (*r* = .07, *p* = .67, *N* = 43). We also correlated the average English comprehensions for each country with the 2019 EF English Proficiency Index (https://www.ef.com/wwen/epi/). Here too there was no significant correlation (*r* = -.18, *p* = .30, *N* = 37). Therefore, our country-level samples may not be representative of the breadth and diversity of each country and the results should be interpreted cautiously and as a basis for generating future hypotheses.

Systematic Review of cross-cultural studies of the Eyes Test

Since our data was based on the English version of the Eyes Test, we wanted to observe if the female advantage was found in translated versions of the Eyes Test. Toward that end, we conducted a systematic review of cross-cultural studies that used translated versions of the Eyes Test. We used the PRISMA model to identify, screen, establish eligibility, and include studies in the review (113). We searched for studies that met the following criteria: (1) published from 2001 (upon first publication of the Eyes Test) until February 2021; (2) included the adult version (36 items) of the Eyes Test (not the child version) that was not modified in any way except for translations into another language (e.g., not shortened, lengthened, or with altered photograph stimuli; (3) included non-clinical samples; (4) analysed sex effects in their sample, including those that did find sex differences and those that did not; (5) *n* > 20 for each sex; and (6) was published in English. We first searched Pubmed for the terms: "Reading the Mind in the Eyes", "Eyes Test", and "RMET". We then searched Google Scholar for the same terms, and added an additional search term, "translate/d" to identify additional studies that administered a translated version of the Eyes Test. The literature search was conducted from January 2021 to March 2021. We excluded studies that were only abstracts, conference proceedings, or grey literature. From each study, we extracted the study meta data, reported means, standard deviations, and results from significant testing for each study that met the inclusion criteria. The resulted PRISMA diagram is presented in **Fig. S1**.

References

1. S. Baron-Cohen, *Mindblindness: An essay on autism and theory of mind* (1995) https:/doi.org/10.1027//0269-8803.13.1.57.
2. S. Baron-Cohen, "Theory of mind and autism: A fifteen year review." in *Understanding Other Minds*, (2000).
3. S. Baron-Cohen, "Theory of mind and autism: A review" in (2004) https:/doi.org/10.1016/s0074-7750(00)80010-5.
4. T. A. Russell, U. Schmidt, L. Doherty, V. Young, K. Tchanturia, Aspects of social cognition in anorexia nervosa: Affective and cognitive theory of mind. *Psychiatry Research* (2009) https:/doi.org/10.1016/j.psychres.2008.10.028.
5. D. de Achával, *et al.*, Emotion processing and theory of mind in schizophrenia patients and their unaffected first-degree relatives. *Neuropsychologia* (2010) https:/doi.org/10.1016/j.neuropsychologia.2009.12.019.
6. J. Decety, K. J. Michalska, Y. Akitsuki, B. B. Lahey, Atypical empathic responses in adolescents with aggressive conduct disorder: A functional MRI investigation. *Biological Psychology* (2009) https:/doi.org/10.1016/j.biopsycho.2008.09.004.
7. E. A. Fertuck, *et al.*, Enhanced reading the mind in the eyes in borderline personality disorder compared to healthy controls. *Psychological Medicine* (2009) https:/doi.org/10.1017/S003329170900600X.

8. C. Campos, *et al.*, Refining the link between psychopathy, antisocial behavior, and empathy: A meta-analytical approach across different conceptual frameworks. *Clinical Psychology Review* **94** (2022).
9. S. R. Beck, Interaction between comparative psychology and cognitive development. *Current Opinion in Behavioral Sciences* (2017) https:/doi.org/10.1016/j.cobeha.2017.07.002.
10. V. E. Stone, S. Baron-Cohen, A. Calder, J. Keane, A. Young, Acquired theory of mind impairments in individuals with bilateral amygdala lesions. *Neuropsychologia* **41** (2003).
11. O. Dal Monte, *et al.*, The left inferior frontal gyrus is crucial for reading the mind in the eyes: Brain lesion evidence. *Cortex* **58** (2014).
12. V. Diveica, K. Koldewyn, R. J. Binney, Establishing a role of the semantic control network in social cognitive processing: A meta-analysis of functional neuroimaging studies. *Neuroimage* **245** (2021).
13. L. Kulke, M. Reiß, H. Krist, H. Rakoczy, How robust are anticipatory looking measures of Theory of Mind? Replication attempts across the life span. *Cognitive Development* **46** (2018).
14. S. Baron Cohen, "Precursors to a theory of mind: Understanding attention in others" in *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*, (1991).
15. K. K. Oniski, R. Baillargeon, Do 15-month-old infants understand false beliefs? *Science (1979)* (2005) https:/doi.org/10.1126/science.1107621.
16. A. M. Leslie, Pretense and Representation: The Origins of "Theory of Mind." *Psychological Review* (1987) https:/doi.org/10.1037/0033-295X.94.4.412.
17. M. Köster, X. Ohmer, T. D. Nguyen, J. Kärtner, Infants Understand Others' Needs. *Psychological Science* **27** (2016).
18. S. Baron-Cohen, *Mindblindness: An essay on autism and theory of mind* (1995) https:/doi.org/10.1027//0269-8803.13.1.57.
19. G. Aschersleben, T. Hofer, B. Jovanovic, The link between infant attention to goal-directed action and later theory of mind abilities. *Developmental Science* (2008) https:/doi.org/10.1111/j.1467-7687.2008.00736.x.
20. S. A. Miller, Children's Understanding of Second-Order Mental States. *Psychological Bulletin* (2009) https:/doi.org/10.1037/a0016854.
21. V. E. Stone, S. Baron-Cohen, R. T. Knight, Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience* (1998) https:/doi.org/10.1162/089892998562942.
22. S. Baron-Cohen, S. Wheelwright, The empathy quotient: An investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders* (2004) https:/doi.org/10.1023/B:JADD.0000022607.19833.00.
23. I. Dumontheil, I. A. Apperly, S. J. Blakemore, Online usage of theory of mind continues to develop in late adolescence. *Developmental Science* (2010) https:/doi.org/10.1111/j.1467-7687.2009.00888.x.
24. H. C. Barrett, *et al.*, Early false-belief understanding in traditional non-Western societies. *Proceedings of the Royal Society B: Biological Sciences* **280** (2013).
25. A. Shahaeian, C. C. Peterson, V. Slaughter, H. M. Wellman, Culture and the Sequence of Steps in Theory of Mind Development. *Developmental Psychology* **47** (2011).

26. Z. Boraston, S. J. Blakemore, R. Chilvers, D. Skuse, Impaired sadness recognition is linked to social interaction deficit in autism. *Neuropsychologia* (2007) https:/doi.org/10.1016/j.neuropsychologia.2006.11.010.
27. H. Wimmer, J. Perner, Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* (1983) https:/doi.org/10.1016/0010-0277(83)90004-5.
28. S. Baron-Cohen, S. Wheelwright, J. Hill, Y. Raste, I. Plumb, The ' ' Reading the Mind in the Eyes ' ' Test Revised Version : A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *J. Child Psychol. Psychiat. Association for Child Psychology and Psychiatry* (2001) https:/doi.org/10.1111/1469-7610.00715.
29. S. Baron-Cohen, *et al.*, The "reading the mind in the eyes" test: Complete absence of typical sex difference in 400 men and women with autism. *PLoS ONE* (2015) https:/doi.org/10.1371/journal.pone.0136521.
30. P. Maurage, *et al.*, The "Reading the Mind in the Eyes" test as a new way to explore complex emotions decoding in alcohol dependence. *Psychiatry Research* (2011) https:/doi.org/10.1016/j.psychres.2011.06.015.
31. C. Heitz, *et al.*, Cognitive and affective theory of mind in dementia with Lewy bodies and Alzheimer's disease. *Alzheimer's Research and Therapy* (2016) https:/doi.org/10.1186/s13195-016-0179-9.
32. S. Baron-Cohen, *et al.*, Social intelligence in the normal and autistic brain: An fMRI study. *European Journal of Neuroscience* **11** (1999).
33. R. J. Holt, *et al.*, "Reading the Mind in the Eyes": An fMRI study of adolescents with autism and their siblings. *Psychological Medicine* **44** (2014).
34. V. Warrier, *et al.*, Genome-wide meta-analysis of cognitive empathy: Heritability, and correlates with sex, neuropsychiatric conditions and cognition. *Molecular Psychiatry* (2018) https:/doi.org/10.1038/mp.2017.122.
35. E. Chapman, *et al.*, Fetal testosterone and empathy: evidence from the Empathy Quotient (EQ) and the "Reading the Mind in the Eyes" test. *Soc Neurosci* **1**, 135–48 (2006).
36. G. Domes, M. Heinrichs, A. Michel, C. Berger, S. C. Herpertz, Oxytocin Improves "Mind-Reading" in Humans. *Biological Psychiatry* (2007) https:/doi.org/10.1016/j.biopsych.2006.07.015.
37. V. Warrier, V. Chee, P. Smith, B. Chakrabarti, S. Baron-Cohen, A comprehensive meta-analysis of common genetic variants in autism spectrum conditions. *Molecular Autism* (2015) https:/doi.org/10.1186/s13229-015-0041-0.
38. E. Chapman, *et al.*, Fetal testosterone and empathy: evidence from the empathy quotient (EQ) and the "reading the mind in the eyes" test. *Soc Neurosci* (2006) https:/doi.org/10.1080/17470910600992239.
39. F. Uzefovsky, *et al.*, The oxytocin receptor gene predicts brain activity during an emotion recognition task in autism. *Molecular Autism* (2019) https:/doi.org/10.1186/s13229-019-0258-4.
40. C. A. Baker, E. Peterson, S. Pulos, R. A. Kirkland, Eyes and IQ: A meta-analysis of the relationship between intelligence and "Reading the Mind in the Eyes." *Intelligence* (2014) https:/doi.org/10.1016/j.intell.2014.03.001.
41. S. L. K. Stewart, J. A. Kirkham, Predictors of individual differences in emerging adult theory of mind. *Emerging Adulthood*, 2167696820926300 (2020).

42. R. A. Kirkland, E. Peterson, C. A. Baker, S. Miller, S. Pulos, Meta-analysis reveals adult female superiority in "Reading the Mind in the Eyes Test." *North American Journal of Psychology* (2013) https:/doi.org/http://dx.doi.org/10.1108/17506200710779521.
43. V. Warrier, *et al.*, Genome-wide meta-analysis of cognitive empathy: heritability, and correlates with sex, neuropsychiatric conditions and cognition. *Molecular Psychiatry* (2017).
44. J. P. Hill, M. E. Lynch, "The Intensification of Gender-Related Role Expectations during Early Adolescence" in *Girls at Puberty*, (1983) https:/doi.org/10.1007/978-1-4899-0354-9_10.
45. W. Wood, A. H. Eagly, "Biosocial Construction of Sex Differences and Similarities in Behavior" in *Advances in Experimental Social Psychology*, (2012).
46. T. A. Russell, K. Tchanturia, Q. Rahman, U. Schmidt, Sex differences in theory of mind: A male advantage on Happé's "cartoon" task. *Cognition and Emotion* (2007) https:/doi.org/10.1080/02699930601117096.
47. F. Quesque, Y. Rossetti, What Do Theory-of-Mind Tasks Actually Measure? Theory and Practice. *Perspectives on Psychological Science* **15** (2020).
48. S. Olderbak, O. Wilhelm, A. Hildebrandt, J. Quoidbach, Sex differences in facial emotion perception ability across the lifespan. *Cognition and Emotion* **33** (2019).
49. R. C. Gur, *et al.*, Age group and sex differences in performance on a computerized neurocognitive battery in children age 8-21. *Neuropsychology* **26** (2012).
50. V. Warrier, S. Baron-Cohen, Genetic contribution to "theory of mind" in adolescence. *Scientific Reports* (2018) https:/doi.org/10.1038/s41598-018-21737-8.
51. N. J. Sasson, *et al.*, Controlling for Response Biases Clarifies Sex and Age Differences in Facial Affect Recognition. *Journal of Nonverbal Behavior* **34** (2010).
52. R. Cabello, M. A. Sorrel, I. Fernández-Pinto, N. Extremera, P. Fernández-Berrocal, Age and gender differences in ability emotional intelligence in adults: A cross-sectional study. *Developmental Psychology* **52** (2016).
53. A. Megías-Robles, *et al.*, The 'Reading the mind in the Eyes'' test and emotional intelligence.' *Royal Society Open Science* **7** (2020).
54. E. mac Giolla, P. J. Kajonius, Sex differences in personality are larger in gender equal countries: Replicating and extending a surprising finding. *International Journal of Psychology* (2018) https:/doi.org/10.1002/ijop.12529.
55. G. Stoet, D. C. Geary, The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education. *Psychological Science* (2018) https:/doi.org/10.1177/0956797617741719.
56. A. Falk, J. Hermle, Relationship of gender differences in preferences to economic development and gender equality. *Science (1979)* (2018) https:/doi.org/10.1126/science.aas9899.
57. P. T. Costa, A. Terracciano, R. R. McCrae, Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology* (2001) https:/doi.org/10.1037/0022-3514.81.2.322.
58. R. A. Lippa, Sex differences in personality traits and gender-related occupational preferences across 53 nations: Testing evolutionary and social-environmental theories. *Archives of Sexual Behavior* **39** (2010).

59. D. P. Schmitt, A. Realo, M. Voracek, J. Allik, Why Can't a Man Be More Like a Woman? Sex Differences in Big Five Personality Traits Across 55 Cultures. *Journal of Personality and Social Psychology* (2008) https:/doi.org/10.1037/0022-3514.94.1.168.
60. R. Su, J. Rounds, P. I. Armstrong, Men and Things, Women and People: A Meta-Analysis of Sex Differences in Interests. *Psychological Bulletin* **135** (2009).
61. A. Herlitz, J. Lovén, Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition* **21** (2013).
62. G. Kreutz, E. Schubert, L. A. Mitchell, Cognitive styles of music listening. *Music Perception* (2008) https:/doi.org/10.1525/mp.2008.26.1.57.
63. Simon. Baron-Cohen, *et al.*, Why are autism spectrum conditions more prevalent in males? *PLoS Biol* **9**, e1001081 (2011).
64. M. Cabinio, *et al.*, Mind-reading ability and structural connectivity changes in aging. *Frontiers in Psychology* (2015) https:/doi.org/10.3389/fpsyg.2015.01808.
65. J. D. Henry, L. H. Phillips, T. Ruffman, P. E. Bailey, A meta-analytic review of age differences in theory of mind. *Psychology and Aging* (2013) https:/doi.org/10.1037/a0030677.
66. G. Slessor, L. H. Phillips, R. Bull, Exploring the Specificity of Age-Related Differences in Theory of Mind Tasks. *Psychology and Aging* (2007) https:/doi.org/10.1037/0882-7974.22.3.639.
67. D. Dodell-Feder, K. J. Ressler, L. T. Germine, Social cognition or social class and culture? on the interpretation of differences in social cognitive performance. *Psychological Medicine* (2018) https:/doi.org/10.1017/S003329171800404X.
68. L. R. Goldberg, The Development of Markers for the Big-Five Factor Structure. *Psychological Assessment* (1992) https:/doi.org/10.1037/1040-3590.4.1.26.
69. D. M. Greenberg, V. Warrier, C. Allison, S. Baron-Cohen, Testing the Empathizing-Systemizing theory of sex differences and the Extreme Male Brain theory of autism in half a million people. *Proc Natl Acad Sci U S A*, 201811032 (2018).
70. S. Wheelwright, *et al.*, Predicting Autism Spectrum Quotient (AQ) from the Systemizing Quotient-Revised (SQ-R) and Empathy Quotient (EQ). *Brain Research* (2006) https:/doi.org/10.1016/j.brainres.2006.01.012.
71. H. Takeuchi, *et al.*, Empathizing associates with mean diffusivity. *Scientific Reports* **9** (2019).
72. H. Takeuchi, *et al.*, Association between resting-state functional connectivity and empathizing/systemizing. *Neuroimage* **99** (2014).
73. H. Takeuchi, *et al.*, Regional gray matter volume is associated with empathizing and systemizing in young adults. *PLoS ONE* **9** (2014).
74. H. Takeuchi, *et al.*, White matter structures associated with empathizing and systemizing in young adults. *Neuroimage* **77** (2013).
75. A. Kobayashi, *et al.*, Increased grey matter volume of the right superior temporal gyrus in healthy children with autistic cognitive style: A VBM study. *Brain and Cognition* **139** (2020).
76. M.-C. Lai, *et al.*, Individual differences in brain structure underpin empathizing-systemizing cognitive styles in male adults. *Neuroimage* **61**, 1347–54 (2012).
77. F. Focquaert, M. S. Steven-Wheeler, S. Vanneste, K. W. Doron, S. M. Platek, Mindreading in individuals with an empathizing versus systemizing cognitive style: An fMRI study. *Brain Res Bull* **83**, 214–22 (2010).

78. P. J. Rentfrow, *et al.*, Divided we stand: Three psychological regions of the united states and their political, economic, social, and health correlates. *Journal of Personality and Social Psychology* (2013) https:/doi.org/10.1037/a0034434.

79. P. J. Rentfrow, M. Jokela, M. E. Lamb, Regional personality differences in Great Britain. *PLoS ONE* (2015) https:/doi.org/10.1371/journal.pone.0122245.

80. P. C. Austin, J. E. Hux, A brief note on overlapping confidence intervals. *Journal of Vascular Surgery* **36** (2002).

81. P. J. Rentfrow, *et al.*, Divided we stand: Three psychological regions of the united states and their political, economic, social, and health correlates. *Journal of Personality and Social Psychology* (2013) https:/doi.org/10.1037/a0034434.

82. C. J. Hill, H. S. Bloom, A. R. Black, M. W. Lipsey, Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives* **2** (2008).

83. M. del Giudice, Measuring Sex Differences and Similarities. *Gender and sexuality development: Contemporary theory and research* (2019).

84. J. Cohen, Statistical power analysis for the behavioural sciences. Hillside. *NJ: Lawrence Earlbaum Associates* (1988) https:/doi.org/10.1111/1467-8721.ep10768783.

85. D. C. Funder, D. J. Ozer, Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science* (2019) https:/doi.org/10.1177/2515245919847202.

86. W. Wood, A. H. Eagly, A cross-cultural analysis of the behavior of women and men: Implications for the origins of sex differences. *Psychological Bulletin* (2002) https:/doi.org/10.1037/0033-2909.128.5.699.

87. F. Rahman, *et al.*, Sources of Cognitive Conflict and Their Relevance to Theory-of-Mind Proficiency in Healthy Aging: A Preregistered Study. *Psychological Science* **32** (2021).

88. M. Pardini, P. F. Nichelli, Age-related decline in mentalizing skills across adult life span. *Experimental Aging Research* **35** (2009).

89. K. A. Muscatell, *et al.*, Social status modulates neural activity in the mentalizing network. *Neuroimage* (2012) https:/doi.org/10.1016/j.neuroimage.2012.01.080.

90. M. W. Kraus, S. Côté, D. Keltner, Social Class, Contextualism, and Empathic Accuracy. *Psychological Science* (2010) https:/doi.org/10.1177/0956797610387613.

91. M. W. Kraus, D. Keltner, Signs of socioeconomic status: A thin-slicing approach. *Psychological Science* (2009) https:/doi.org/10.1111/j.1467-9280.2008.02251.x.

92. P. K. Piff, M. W. Kraus, S. Côté, B. H. Cheng, D. Keltner, Having Less, Giving More: The Influence of Social Class on Prosocial Behavior. *Journal of Personality and Social Psychology* (2010) https:/doi.org/10.1037/a0020092.

93. Y. Gorodnichenko, G. Roland, Individualism, innovation, and long-run growth. *Proc Natl Acad Sci U S A* (2011) https:/doi.org/10.1073/pnas.1101933108.

94. G. Handley, J. T. Kubota, T. Li, J. Cloutier, Black “Reading the Mind in the Eyes” task: The development of a task assessing mentalizing from black faces. *PLoS ONE* (2019) https:/doi.org/10.1371/journal.pone.0221867.

95. K. Lee, P. C. Quinn, O. Pascalis, Face Race Processing and Racial Bias in Early Development: A Perceptual-Social Linkage. *Current Directions in Psychological Science* **26** (2017).

96. E. G. Fernández-Abascal, R. Cabello, P. Fernández-Berrocal, S. Baron-Cohen, Test-retest reliability of the “Reading the Mind in the Eyes” test: A one-year follow-up study. *Molecular Autism* (2013) https:/doi.org/10.1186/2040-2392-4-33.

97. M. U. Hallerbäck, T. Lugnegård, F. Hjärthag, C. Gillberg, The Reading the Mind in the eyes test: Test-retest reliability of a swedish version. *Cognitive Neuropsychiatry* (2009) https:/doi.org/10.1080/13546800902901518.

98. M. Vellante, *et al.*, The "reading the Mind in the Eyes" test: Systematic review of psychometric properties and a validation study in Italy. *Cognitive Neuropsychiatry* (2013) https:/doi.org/10.1080/13546805.2012.721728.

99. B. S. Khorashad, *et al.*, The "Reading the Mind in the Eyes" Test: Investigation of Psychometric Properties and Test–Retest Reliability of the Persian Version. *Journal of Autism and Developmental Disorders* (2015) https:/doi.org/10.1007/s10803-015-2427-4.

100. M. C. Pfaltz, *et al.*, The Reading the Mind in the Eyes Test: Test- retest Reliability and Preliminary Rsychometric Rroperties of the German Version. *International Journal of Advances in Psychology* (2013).

101. J. H. McDermott, A. F. Schultz, E. A. Undurraga, R. A. Godoy, Indifference to dissonance in native Amazonians reveals cultural variation in music perception. *Nature* (2016) https:/doi.org/10.1038/nature18635.

102. K. Reinecke, K. Z. Gajos, Labin the wild: Conducting large-scale online experiments with uncompensated samples in *CSCW 2015 - Proceedings of the 2015 ACM International Conference on Computer-Supported Cooperative Work and Social Computing*, (2015) https:/doi.org/10.1145/2675133.2675246.

103. S. Baron-Cohen, S. Wheelwright, A. Spong, J. Lawson, Studies of Theory of Mind: Are Intuitive Physics and Intuitive Psychology Independent? *Journal of Developmental and Learning Disorders* (2001).

104. I. Vogindroukas, E. N. Chelas, N. E. Petridis, Reading the mind in the eyes test (Children's Version): A comparison study between children with typical development, children with high-functioning autism and typically developed adults. *Folia Phoniatrica et Logopaedica* (2014) https:/doi.org/10.1159/000363697.

105. A. van der Meulen, S. Roerig, D. de Ruyter, P. van Lier, L. Krabbendam, A comparison of children's ability to read children's and adults' mental states in an adaptation of the reading the mind in the eyes task. *Frontiers in Psychology* (2017) https:/doi.org/10.3389/fpsyg.2017.00594.

106. S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, E. Clubley, The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. *Journal of Autism and Developmental Disorders* (2001) https:/doi.org/10.1023/A:1005653411471.

107. Simon. Baron-Cohen, J. Richler, D. Bisarya, N. Gurunathan, S. J. Wheelwright, The systemizing quotient: an investigation of adults with Asperger syndrome or high-functioning autism, and normal sex differences. *Philos Trans R Soc Lond B Biol Sci* **358**, 361–74 (2003).

108. S. Gosling, P. Rentfrow, W. S. Jr, A Very Brief Measure of the Big Five Personality Domains. *Journal of Research in …* (2003) https:/doi.org/10.1016/S0092-6566(03)00046-1.

109. E. Diener, R. A. Emmons, R. J. Larsem, S. Griffin, The Satisfaction With Life Scale. *Journal of Personality Assessment* (1985) https:/doi.org/10.1207/s15327752jpa4901_13.

110. A. Wakabayashi, *et al.*, Development of short forms of the Empathy Quotient (EQ-Short) and the Systemizing Quotient (SQ-Short). *Personality and Individual Differences* (2006) https:/doi.org/10.1016/j.paid.2006.03.017.

111. S. H. Schwartz, A theory of cultural value orientations: Explication and applications. *Comparative Sociology* (2006) https:/doi.org/10.1163/156913306778667357.

112. S. H. Schwartz, Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology* (1992) https:/doi.org/10.1016/S0065-2601(08)60281-6.

113. D. Moher, *et al.*, Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine* **6** (2009).

Acknowledgements: DMG was funded in part by the Zuckerman STEM Leadership Program. SBC received funding from the Wellcome Trust 214322\Z\18\Z. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The results leading to this publication have received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777394 for the project AIMS-2-TRIALS. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and AUTISM SPEAKS, Autistica, SFARI. SBC also received funding from the Autism Centre of Excellence, SFARI, the Templeton World Charitable Fund, the MRC, and the NIHR Cambridge Biomedical Research Centre. The research was supported by the National Institute for Health Research (NIHR) Applied Research Collaboration East of England. All research at the Department of Psychiatry in the University of Cambridge is supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) and NIHR Applied Research Centre. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. Any views expressed are those of the author(s) and not necessarily those of the funder.

Figure Legends

Fig. 1. Schematic overview of the study.

In this study, we investigated three major questions. Are there on-average sex differences on the Eyes Test; are there on-average age differences on the Eyes Test; and, do the on-average sex differences, if any, appear across countries? These three questions are represented by large bold font. We also asked secondary questions: what are the sociodemographic and cognitive/personality factors associated with scores on the Eyes Test and what are the country-level variables (PESH indicators) associated with country-level sex differences on the Eyes Test? Data used to address the former question (including D-scores, Big Five personality traits, education, income) are represented in smaller non-bold font. PESH indicators are not visualized in this diagram. For each question, the primary dataset was the discovery dataset from Lab in the Wild (blue box). We used three validation datasets to validate and extend the results-validation dataset A from Cambridge Psychology (red box), validation dataset B from Musical Universe (purple box), and validation dataset C from Musical Universe (green box). If an arrow appears to go through/underneath a box, then the variable in the box is not included in the specified dataset of the arrow, which can be discerned by the colour of the arrow.

Fig. 2. Sex differences on the English version of the Eyes Test across the discovery and validation datasets.

Each plot displays the conditional effects of sex (population-level predictor) with 95% credible intervals. As can be seen, there is evidence for an on-average female advantage in each of the four datasets.

Fig. 3. The effect of sex differences on the English version of the Eyes Test in each of 57 countries in the discovery dataset.

Beta values with 95% credible intervals from multi-level Bayesian analysis are plotted for each of 57 countries. Beta values above zero indicate a descriptive female advantage and beta values below zero indicate a male advantage. As can be seen, 36 countries have a lower bound credible interval ≥ 0, indicating a female advantage, while no countries have a higher bound credible interval ≤ 0, which would indicate a male advantage.

Fig. 4 Geographic distribution of ω_t for the English version of the Eyes Test across 57 countries.

This figure displays Omega total (ωt) for each of the 57 countries observed in the discovery dataset. Lightened yellow colours indicate lower values while darker red colours indicate higher values.

Fig. 5. Mean scores on the English version of the Eyes Test by age and sex in the discovery dataset.

This figure visualizes the results from the constrained non-linear regression analysis performed separately for females and males. Results are age and sex differences in Eyes Test scores from 16 to 70 years of age. The figure also identifies inflection points in performance across this age range. Average scores and associated 95% confidence intervals are charted for females and males at each age year.

Fig. 6. **An item from the 'Reading the Mind in the Eyes Test'.**

This photograph in this figure is from item 19 of the Eyes Test. Underneath the photograph are four answer choices: arrogant, grateful, sarcastic, and tentative. The correct answer is tentative.

Tables

Table 1. Sex differences on the Eyes Test in the discovery and validations datasets.

			Eyes Test scores		Sex differences			Reliability	
		N	*M*	*SD*	*Beta*	*SE*	*95% CIs*	ω_t	ω_h
Discovery	females	148,923	27.62	3.92	.17	.00	.16, .18	.80	.50
	males	142,694	26.94	4.05					
Validation A	females	422	26.90	3.50	.23	.08	.07, .39	.76	.45
	males	220	26.03	4.32					
Validation B	females	2,947	14.33	2.02	.27	.03	.22, .32	.74	.38
	males	2,293	13.75	2.19					
Validation C	females	388	28.68	3.33	.19	.08	.03, .34	.75	.32
	males	281	28.02	3.23					

This table provides the sample size for each dataset, along with the mean (M) and standard deviation (SD) of Eyes Test scores for each sex, beta, 95% credible intervals, and both omega total (ω_t) and omega hierarchical (ω_h) reliability coefficients. The maximum possible score on the Eyes Test is 36 for discovery, validation A, and validation C. The total score on the Eyes Test is 18 for validation B.