Psychological Medicine

cambridge.org/psm

Social cognition or social class and culture? On the interpretation of differences in social cognitive performance

David Dodell-Feder[1], Kerry J. Ressler[2,3] and Laura T. Germine[3,4]

[1]Department of Psychology, University of Rochester, Rochester, NY, USA; [2]Division of Depression and Anxiety, McLean Hospital, Belmont, MA, USA; [3]Department of Psychiatry, Harvard Medical School, Boston, MA, USA and [4]Institute for Technology in Psychiatry, McLean Hospital, Belmont, MA, USA

Original Article

Cite this article: Dodell-Feder D, Ressler KJ, Germine LT (2019). Social cognition or social class and culture? On the interpretation of differences in social cognitive performance. *Psychological Medicine* 1–13. https://doi.org/10.1017/S003329171800404X

Received: 6 August 2018
Revised: 19 November 2018
Accepted: 12 December 2018

Key words:
Mental state understanding; reading the mind in the eyes task; research domain criteria; social cognition; theory of mind

Author for correspondence:
David Dodell-Feder,
E-mail: d.dodell-feder@rochester.edu

Abstract

Background. The ability to understand others' mental states carries profound consequences for mental and physical health, making efforts at validly and reliably assessing mental state understanding (MSU) of utmost importance. However, the most widely used and current NIMH-recommended task for assessing MSU – the Reading the Mind in the Eyes Task (RMET) – suffers from potential assessment issues, including reliance on a participant's vocabulary/intelligence and the use of culturally biased stimuli. Here, we evaluate the impact of demographic and sociocultural factors (age, gender, education, ethnicity, race) on the RMET and other social and non-social cognitive tasks in an effort to determine the extent to which the RMET may be unduly influenced by participant characteristics.

Methods. In total, 40 248 international, native-/primarily English-speaking participants between the ages of 10 and 70 completed one of five measures on TestMyBrain.org: RMET, a shortened version of RMET, a multiracial emotion identification task, an emotion discrimination task, and a non-social/non-verbal processing speed task (digit symbol matching).

Results. Contrary to other tasks, performance on the RMET increased across the lifespan. Education, race, and ethnicity explained more variance in RMET performance than the other tasks, and differences between levels of education, race, and ethnicity were more pronounced for the RMET than the other tasks such that more highly educated, non-Hispanic, and White/Caucasian individuals performed best.

Conclusions. These data suggest that the RMET may be unduly influenced by social class and culture, posing a serious challenge to assessing MSU in clinical populations given shared variance between social status and psychiatric illness.

Success in the social world hinges upon our ability to decipher and infer the hidden beliefs, emotions, and intentions of others; a process commonly known as 'theory of mind'. Indeed, a large body of research has demonstrated that our ability to understand others' mental states is associated with a variety of positive social outcomes including increased popularity (Slaughter *et al.*, 2015), improved interpersonal rapport (Blatt *et al.*, 2010; Todd *et al.*, 2011), prosocial behavior (Imuta *et al.*, 2016), positive evaluations of the perspective-taker (Goldstein *et al.*, 2014), and the well-being of perspective-taking recipients (Dodell-Feder *et al.*, 2016), among many other outcomes (Todd and Galinsky, 2014). In contrast, difficulty with mental state understanding and disruption to its neural bases is associated with a variety of negative outcomes, including friendlessness (Fink *et al.*, 2015), social amotivation and isolation (Dodell-Feder *et al.*, 2014*a*, 2014*b*), and risk for severe psychiatric illness, such as schizophrenia-spectrum disorders (Kim *et al.*, 2011). These findings are even more sobering when considering the robust relationships among social isolation, psychiatric illness, and mortality (Holt-Lunstad *et al.*, 2015; Walker *et al.*, 2015). Thus, the health consequences of MSU are not to be ignored.

For these reasons, the importance of valid and reliable assessments of MSU cannot be overstated. Failure to detect MSU impairments could lead researchers and clinicians to overlook etiological factors, fail to identify someone at risk leading them on a path toward mental and physical health decline, or incorrectly conclude that a treatment is efficacious when it is not. On the other hand, detecting impairments when they do not exist could lead researchers and clinicians on a wayward path of research to elucidate a specious causal pathway, misidentify someone as being at-risk for psychopathology leading to stigma and unnecessary and costly interventions, or incorrectly conclude that a treatment is not efficacious when it is. Said simply, inaccurate MSU assessment may carry deleterious consequences for every area of clinical research, including the study of etiology, risk, and treatment.

Do we have valid and reliable assessments of MSU? In light of the assessment issues highlighted above and their potential consequences, this question has appropriately been taken up

CAMBRIDGE UNIVERSITY PRESS

by several initiatives at the National Institute of Mental Health (NIMH; National Advisory Mental Health Council Workgroup on Tasks and Measures for RDoC, 2016; Pinkham *et al.*, 2018) who define MSU as 'The ability to make judgments and/or attributions about the mental state (intentions, beliefs, desires, emotions) of other animate entities that allows one to predict or interpret their behaviors' (National Advisory Mental Health Council Workgroup on Tasks and Measures for RDoC, 2016), placing it at the nexus of theory of mind, social perception (i.e. decoding and interpreting social information), and emotion processing (i.e. perceiving emotions) (Pinkham *et al.*, 2014). The answer to this question, as per the NIMH (National Advisory Mental Health Council Workgroup on Tasks and Measures for RDoC, 2016), is unclear. What is clear is that there are favorite measures among researchers assessing MSU. Likely the most widely used task to assess MSU is the Reading the Mind in the Eyes Task (RMET; Baron-Cohen *et al.*, 2001*a*), which is also the current NIMH recommended task for assessing mental and emotional perspective-taking[1†] as detailed in a 2016 report by the National Advisory Mental Health Council Workgroup on Tasks and Measures for Research Domain Criteria (RDoC; National Advisory Mental Health Council Workgroup on Tasks and Measures for RDoC, 2016). In the RMET, participants view 36 black-and-white photographs, originally selected from magazine articles, of solely the eyes of Caucasian female and male actresses/actors. Participants decide which of four adjectives (e.g. *panicked*, *incredulous*, *despondent*, *interested*) best describes the mental state being expressed in the eyes with the correct answer having been generated through consensus ratings. As of early 2018, the paper detailing the revised version of the test (Baron-Cohen *et al.*, 2001*a*) has been cited close to 1900 times (Web of Science); the RMET has been translated into at least 24 different languages (https://www.autismresearchcentre.com/arc_tests), reflecting its widespread, international use; the RMET is used more than any other measure of MSU in the psychopathology literature (see the following meta-analyses on schizophrenia: Bora *et al.*, 2009; Bora and Pantelis, 2013; bipolar disorder: Bora *et al.*, 2016; and autism: Chung *et al.*, 2014) as well as other fields of study (Dodell-Feder and Tamir, 2018); the RMET is often used in neuroimaging studies of MSU (Molenberghs *et al.*, 2016); and the RMET has been used in clinical trials (Anagnostou *et al.*, 2012). Taken together, well over a decade of research on MSU has largely been based on findings from this task.

What explains the popularity of the RMET? In addition to being quick to administer and highly tolerable by participants (Pinkham *et al.*, 2018), exhibiting sensitivity to clinical impairment, and having its neural bases well-studied, it is often considered to represent an 'advanced test' (Baron-Cohen *et al.*, 2001*b*, p. 241) of MSU as it requires participants to decode nuanced facial expressions and match them to a lexicon of nuanced intrapersonal states. As an ostensible consequence of the task's complexity, it has minimal ceiling effects (Olderbak *et al.*, 2015; although see Black, 2018), which often plague other MSU tasks (see Dodell-Feder *et al.*, 2013 for a brief discussion). That said, inspection of the task reveals some potential issues. For example, task performance is moderately-to-strongly associated with vocabulary (Olderbak *et al.*, 2015), IQ (Baker *et al.*, 2014), and educational attainment (Warrier *et al.*, 2018); perhaps unsurprising given the complex and uncommon vocabulary used as response options (e.g. *aghast*, *tentative*). Additionally, the mental states of the target are actually unknown, and consensus scoring leaves open the possibility that accuracy depends on sharing social norms and beliefs regarding how mental states may be expressed in eyes with the consensus raters (Johnston *et al.*, 2008). In a similar vein, most of the female stimuli depict young women who are wearing cosmetic products, and revealing a limited number of gender normative mental states[2] in a way that may be very different to how women are depicted in other cultures. These concerns are borne out in research revealing sociocultural differences in RMET performance (Adams *et al.*, 2010; Prevost *et al.*, 2014). Of course, the RMET is not unique in these respects. Other MSU tasks suffer from many of the same issues (Corcoran *et al.*, 1995; McDonald *et al.*, 2003; Dodell-Feder *et al.*, 2013) through reliance on verbal ability, the use of racially/ethnically homogenous stimuli, or the depiction of social scenarios in which participants may come to a very different understanding of the characters, not because of poor MSU ability, but because of cultural differences in how they understand the broader social context (see Kohler *et al.*, 2003 for an example in which the issue of stimulus race/ethnicity is appropriately dealt with; see Nisbett, 2004 for a general discussion of cultural differences in cognition). That said, all of these issues are present in the RMET, and the available evidence suggests that performance on this task may be particularly affected by factors that are associated with social class (i.e. education, IQ) and culture (i.e. race, ethnicity).

Here, leveraging four massive web-based datasets of the RMET and other social and non-social cognitive tasks, totaling over 40 000 international, native- or primarily English-speaking participants, we evaluate whether and to what extent performance on social cognitive tasks, and the RMET in particular, is influenced by demographic and sociocultural factors, including age, gender, education, ethnicity, and race. Said otherwise, we ask whether performance on these tasks is contaminated by sociodemographic factors associated with class and culture, among other factors. In doing so, we hope to inform efforts aimed at identifying, creating, and disseminating valid and reliable measures for the assessment of constructs that may be used in future diagnostic systems.

Participants completed one of five measures: (1) the original RMET (RMET-36), (2) a shortened version of the RMET (RMET-16; described in the Methods), (3) a measure of emotion identification requiring participants to label the emotions of multiracial stimuli (henceforth, 'multiracial emotion identification'), (4) a measure of emotion discrimination requiring participants to make *same*/*different* judgments regarding the emotions expressed in Caucasian faces, and (5) digit symbol matching, which is a non-social, non-verbal measure of processing speed tapping visual working memory and sensorimotor speed.

To understand how task characteristics might impact the relationship between sociocultural factors and MSU performance, we selected four tasks that differed in stimulus and task parameters, including racial diversity of faces and reliance on vocabulary. This allowed us to test several possibilities regarding the influence of sociocultural factors on MSU performance (Table 1); that is, whether the relations between sociocultural factors and measure performance is the result of the measure being (a) social (RMET, multiracial emotion identification, emotion discrimination) *v.* non-social (digit symbol matching[3]), (b) requiring emotion labeling (RMET, multiracial emotion identification) *v.* no emotion labeling (emotion discrimination, digit symbol matching), (c) using uncommon vocabulary (RMET) *v.* common vocabulary/no vocabulary

†The notes appear after the main text.

https://doi.org/10.1017/S003329171800404X

Table 1. Measure characteristics and hypotheses

	Characteristic			
	Mental state understanding	Emotion labeling	Uncommon vocabulary	Caucasian-only faces
RMET	✓	✓	✓	✓
Emotion identification	✓	✓		
Emotion discrimination	✓			✓
Digit symbol matching				
	Hypothesis: If the relation between sociocultural factors and measure performance is the result of the effect described in the column heading, we would expect to see the following similarities/differences in sociocultural-performance relations…			
	Mental state understanding effect	Emotion labeling effect	Uncommon vocabulary effect (*class effect*)	Caucasian-only faces effect (*culture effect*)
RMET	+	+	+	+
Emotion identification	+	+	–	–
Emotion discrimination	+	–	–	+
Digit symbol matching	–	–	–	–

A check-mark indicates the presence of the column characteristic. Shared + /– symbols indicates that the relation between sociocultural factors and measure performance is expected to be similar if the column hypothesis is supported.

(multiracial emotion identification, emotion discrimination, digit symbol matching)[4], and (d) using Caucasian-only (RMET, emotion discrimination) *v.* non-Caucasian-only or no face stimuli (multiracial emotion identification, digit symbol matching).

To evaluate the aforementioned possibilities, we compared effect size relationships between task performance and different demographic and sociocultural characteristics across tasks, including age, gender, education, ethnicity, and race.

Methods

Participants

Participants were an international sample of 40 248 native or primarily English-speaking Internet-users between the ages of 10 and 70 ($M = 29.5$, S.D. = 14.36; 57.3% female) who visited the non-profit research initiative website TestMyBrain.org between 2010 and 2017 (Table 1). Most participants were from countries in which the majority of the population is English-speaking and White/European (i.e. USA, Canada, Great Britain, Ireland, Australia, New Zealand, $n = 31310$, 77.8%; $n = 6085$, 15.1% were from non-predominantly White/English-speaking countries; no data were available for $n = 2853$, 7.1% of participants).

No explicit recruitment was performed. Instead, participants come to TestMyBrain.org through search engine results, social-networking websites, or by word of mouth. As such, the participants represent a non-random sample leaving open the possibility of self-selection effects. However, prior research using TestMyBrain.org has shown that the data are reliable and comparable in terms of quality to data collected in the laboratory (Germine *et al.*, 2012), and mirror findings from nationally representative, population-based samples (Hartshorne and Germine, 2015), suggesting that the potential impact of self-selection effects is likely to be minor.

Prior to completing one of the measures described below, participants provided informed consent/assent by electronically signing a form in a manner approved by the Harvard University Committee on the Use of Human Subjects in Research. Since the public nature of the research platform means that requirements of parental consent cannot be validated, and given concerns that any additional requirements related to age may lead to false self-reported age (e.g. giving an age that permits full participation or reduces participant burden; Boyd *et al.*, 2011), the protocol was designed such that participants giving an age <18 were directed to measures that were deemed to be minimal risk for minors and otherwise not required to obtain parental consent (e.g. the measures used in the current study). This consent procedure has been in place since 2009 with no adverse events reported. After completing one of the measures, participants voluntarily provided demographic and sociocultural information including age, gender, education, ethnicity, and race. No personal identifying information was collected in order to avoid social desirability responding. A subset of the participants included in the current study have been included in other published studies (Germine and Hooker, 2011; Hartshorne and Germine, 2015).

Measures

Reading the Mind in the Eyes Task

In the RMET (RMET-36), participants are presented with 36 pictures of solely the eye region from white actors and actresses taken from magazine photographs. Participants are instructed to select which of four adjectives best describes what the person in the picture is thinking or feeling. Each item is scored as *correct* or *incorrect*; thus, scores can range from 0 to 36. The RMET's psychometric properties have been well-studied. On reliability, internal consistency is typically poor, although test-retest

https://doi.org/10.1017/S003329171800404X

reliability estimates are generally acceptable (Olderbak *et al.*, 2015). On validity, RMET scores tend to correlate with scores on other measures of social cognition (Olderbak *et al.*, 2015), differentiates clinical groups with well-documented social cognitive impairments (e.g. schizophrenia, autism) from healthy participants (Chung *et al.*, 2014), and is predictive of functioning in clinical samples (Pinkham *et al.*, 2018).

One possibility is that prior findings relating RMET performance to sociocultural factors is the result of task heterogeneity; that is, while performance for most stimuli load onto a single MSU-related factor that is relatively unbiased, there may also exist low-quality items that tap more into education and sociocultural biases (e.g. items with complicated vocabulary words), which might also account for the relatively low internal reliability of the task. Based on analyses by Olderbak *et al.* (2015), we took a subset of 16 items from the RMET that load most strongly onto a single common factor which we might assume *a priori* indices MSU performance, and had part of our sample complete this 16-item version of the RMET (RMET-16), which represents a more homogeneous stimulus set.

Multiracial Emotion Identification Task

The multiracial emotion matching task is an emotion identification task where the participant has to indicate whether each of a set of 48 faces is happy, sad, angry, or fearful. Faces represent a broad range of adult ages and race/ethnicities, with approximately equal proportions of men and women.

All tasks have been used in other studies and are described elsewhere, except for the Multiracial Emotion Identification Task. To develop this task, we recruited actors from across a range of ages and races/ethnicities, from the Boston Company One theater, as part of the Act Out for Brain Health project. Boston Company One theater has a mission of engaging the city's diverse communities, with an emphasis on diverse actors. Images were taken from video clips of actors portraying different emotions. An initial set of 146 images were selected portraying anger, fear, happiness, sadness, and neutral facial expressions to create an item bank. Images were drawn from this item bank and data were collected from a development sample of $N = 8309$ participants who each saw a subset of 37–53 images. Ultimately, the neutral condition was dropped as these faces were judged with significantly ($ps < 0.01$) poorer reliability than anger, fear, sadness, and happiness (average correlation with rest of items for each emotion category: anger: $r = 0.3$; fear: $r = 0.26$; sadness: $r = 0.2$; happiness: $r = 0.25$; neutral: $r = 0.06$). The reliability of judgments of other emotions did not significantly differ from each other ($ps > 0.1$). The final test includes 48 images that were selected to capture (1) images with consistent judgments of a single emotion, (2) varying levels of difficulty for each emotion, and (3) items with high correlations with overall emotion recognition accuracy to maximize reliability, while preserving the diversity of actors and faces.

Emotion Discrimination Task

We assessed emotion discrimination using the same stimuli and procedure as in prior studies of emotion discrimination (Pitcher *et al.*, 2008; Garrido *et al.*, 2009; Germine *et al.*, 2011; Germine and Hooker, 2011). Briefly, stimuli consisted of grayscale and cropped pictures of six white female Ekman faces expressing either happiness, sadness, surprise, fear, anger, or disgust. Pairs of faces were presented sequentially for 500 ms per face with a 500 ms inter-stimulus interval. Participants were given up to 7 s to indicate whether the faces were expressing the same or a different emotion. There were an equal amount of face pairs depicting the same and different emotions. Prior work using variants of this task has demonstrated that performance on this task selectively taps emotion discrimination *per se*, and not more general aspects of face processing, such as identity discrimination (Pitcher *et al.*, 2008). Furthermore, behavioral and neural response to this task has been shown to track with psychosis vulnerability (Germine *et al.*, 2011; Germine and Hooker, 2011).

Digit Symbol Matching

The Digit Symbol Matching Task is modeled on the Digit Symbol Coding/Substitution tests from the Wechsler Adult Intelligence Scales. In this task, participants are presented with a key where numbers are paired with symbols. The participant is then shown a single symbol and has to indicate which number goes with that symbol. Scores indicate the number of symbols a participant correctly matches in 90 s. This is a measure of processing speed that taps visual working memory and sensorimotor speed.

Data analysis

Data were analyzed in R (R Core Team, 2017). Participants who did not report relevant demographic/sociocultural information were excluded on an analysis-by-analysis basis. Extreme values were transformed using a 95% Winsorization and *z*-scored to facilitate cross-task comparisons. Findings were considered statistically significant at $p < 0.05$ (two-sided) with correction for multiple comparisons (i.e. post-hoc Tukey HSD for simple effects and Bonferroni-adjusted p values for omnibus ANOVAs and t tests). All analyses are accompanied by relevant effect sizes for group differences (Cohen's d and the Common Language Effect Size (CLES) which denotes the probability that a randomly selected score from one group will be larger than a randomly selected score from another group) and variance accounted for (adjusted R^2, η^2).

The relation between age and task performance was analyzed using segmented regression (Muggeo, 2003, 2008). In segmented regression, multiple linear segments are used to model non-linear changes between two variables. This analysis allows for an estimation of *breakpoints* (i.e. ages in which the relation between age and performance changes), and rates of change (i.e. the slope of the linear segments) before and after the breakpoint. Using AIC and BIC values to evaluate model fit, first, we confirmed that the relation between age and performance was non-linear (which was true for all measures) by comparing linear models to two segment models, and then iteratively tested models with an additional segment until model fit did not improve (i.e. AIC/BIC values did not decrease). In cases where the differences in AIC and BIC values between models were discrepant, we report the more conservative, parsimonious model with fewer segments. All other variables were categorical and group differences were assessed with Welch's t tests or ANOVAs and post-hoc Tukey HSD tests. Since we found that for some measures, the effect of gender and age were moderate-to-large, and education, race, and ethnicity in our dataset were non-independent, we conducted an additional set of analyses for each factor controlling for all other factors (including the non-linear effect of age). To simplify these simultaneous regression models, we dichotomized education into high (grad, college, some college) and low (high, middle) categories and race into European/White and non-European/non-White categories. To evaluate cross-task differences among

https://doi.org/10.1017/S003329171800404X

the levels of each factor, we examined and reported effect sizes and their 95% CIs.

Data availability

The data analyzed during the current study are available on the Open Science Framework repository at https://osf.io/tn9vb/.

Results

Participant characteristics and descriptive statistics from each task are presented in Table 2. Statistics are reported in full in the online Supplementary Materials.

Age

We used segmented regression to evaluate performance over the lifespan. This method allows for an estimation of breakpoints and their 95% CIs, which are ages at which the relation between age and performance changes. Lifespan changes in digit symbol matching were best fit by a three-segment model in which performance steeply rose, $b = 0.152$, 95% CI 0.136–0.168, until age 17.16 years, 95% CI 16.82–17.51, when performance peaked over the lifespan. After this age, performance began decreasing, $b = -0.015$, 95% CI −0.019 to −0.011, and at age 35.29 years, 95% CI 32.64–37.94, began decreasing at a significantly greater rate, $b = -0.040$, 95% CI −0.043 to −0.037. These lifespan changes replicate prior reports of lifespan changes in cognition (Salthouse, 2004), confirming that our sampling approach captured individual differences in performance in a way that replicates traditional studies. On the social cognitive tasks, performance across the lifespan was distinct from digit symbol matching, but largely similar between the tasks with one potentially important difference. In all social tasks, performance rose steeply from pre-adolescence to early adulthood (Fig. 1*a*) with the first breakpoint occurring for multiracial emotion identification at age 16.43, 95% CI 15.29–17.57, and the last breakpoint occurring for emotion discrimination at age 21.85 (95% CI 20.31–23.38), with a breakpoint for RMET occurring in between these ages. For multiracial emotion identification and emotion discrimination, the age of these breakpoints represents peak performance with performance leveling off after these breakpoints, and exhibiting no significant change in either direction across the rest of the lifespan, multiracial emotion identification second slope, $b < 0.001$, 95% CI −0.002 to 0.002, emotion discrimination second slope $b = -0.002$, 95% CI −0.005 to 0.001. However, RMET performance continues to rise after its breakpoint, though at a slower rate than in adolescence, RMET-36 second slope $b = 0.006$, 95% CI 0.005–0.008, RMET-16 second slope $b = 0.005$, 95% CI 0.002–0.007. Results were unchanged when controlling for gender.

To summarize, performance on all measures exhibits a steep rate of improvement through adolescence, reaching a relatively similar breakpoint in late adolescence/early adulthood, between the ages of approximately 16 and 22. For all measures except for RMET, this is where performance peaks, after which performance either significantly declines, as with digit symbol matching, or remains stable, as with multiracial emotion identification and emotion discrimination. In contrast, performance on RMET continues to increase with age; an effect only typically observed with crystallized cognitive abilities, such as vocabulary or knowledge, which increase over the lifespan (Hartshorne and Germine, 2015). This suggests that in contrast to other social cognitive skills, in which performance peaks in late adolescence/early adulthood, performance on RMET may reflect the contribution on vocabulary or some other variable important to complex MSU, which continues to increase over the lifespan.

Gender

Females outperformed males on all social tasks, $ts > 5.43$, $ps < 0.001$, with comparable, small, effect sizes ranging from $d = 0.16$, 95% CI 0.10–0.22, CLES = 54.48% on multiracial emotion identification to $d = 0.26$, 95% CI 0.21–0.31, CLES = 57.36% on RMET-16 (Fig. 1*b*). The similar direction and magnitude of these effects across social cognitive tasks suggests a robustness to the female advantage in social cognition. However, female advantage was specific to social cognitive tasks. On digit symbol matching, males outperformed females, $t_{(15\,540)} = -7.08$, $p < 0.001$, with the difference being smaller in magnitude than the gender difference in social cognition, $d = -0.11$, 95% CI −0.14 to −0.08, CLES = 46.81%.

Education

There was a clear education effect across all tasks, individual task ANOVA $Fs > 8.97$, $ps < 0.001$, such that participants with higher education generally outperformed those with less education, with the biggest differences occurring between those with the highest levels of education (graduate school, college) and lowest levels (middle school). That said, education effects varied substantially across tasks. Education explained 1.35 times more variance in RMET-36 performance, $\eta^2 = 0.0597$, 95% CI 0.0501–0.06914, than the other racially homogenous measure, emotion discrimination $\eta^2 = 0.0441$, 95% CI 0.0311–0.0571, 7.96 times more variance than the other measure which used multiracial stimuli, multiracial emotion identification $\eta^2 = .0075$, 95% CI 0.0029–0.0125, and 6.78 times more variance than the non-social measure, digit symbol coding $\eta^2 = 0.0088$, 95% CI 0.0058–0.0119 (Fig. 2*a*). Moreover, group differences were more reliable and biggest in magnitude for both versions of the RMET (RMET-36 and RMET-16: all post-hoc differences were significant except graduate *v.* college education), with effect sizes ranging from $d = 0.04$–1.05, CLES = 51.12–77.14% for RMET-36 and $d = -0.01$[5]–0.81, CLES = 49.68–71.59% for RMET-16. In contrast, more highly educated participants less consistently outperformed less educated participants for emotion discrimination, multiracial emotion identification, and digit symbol matching, with effect sizes in the small-to-large range for emotion discrimination, range $d = 0.09$–0.72, CLES = 52.41–69.38%, and small-to-moderate range for multiracial emotion identification, range $d = -0.05$–0.39, CLES = 48.56–60.80% and digit symbol matching, range $d = -0.17$–0.39, CLES = 45.09–60.98%. Together, the pattern of results suggests that while better education is associated with better performance across cognitive and social cognitive measures, this effect is magnified with the RMET (i.e. the magnitude of education-level differences on performance is generally larger with the RMET *v.* the other measures). Furthermore, we observed comparable education effects on digit symbol coding and multiracial emotion identification, the social cognitive task that uses multiracial/multiethnic stimuli. This suggests that the effect of education on social cognitive performance can be mitigated with culturally diverse stimuli and/or reduced vocabulary demands.

https://doi.org/10.1017/S003329171800404X

Table 2. Participant characteristics and task performance

Measure	Factor	Level	*N* (%)	*M*	s.d.	Range	α
RMET-36[a]			9271	74.05	12.1	41.7–94.4	0.71
	Country						
		Majority English/White	7408 (79.9)				
		Non-majority English/White	798 (8.6)				
		No data	1065 (11.5)				
	Age			29.8	14.5	10–70	
	Gender						
		Female	5732 (62.8)				
		Male	3401 (37.2)				
	Education						
		Middle	357 (4.1)				
		High	2325 (26.8)				
		Some college	2544 (29.3)				
		College	1898 (21.8)				
		Graduate	1566 (18.0)				
	Ethnicity						
		Non-Hispanic	7909 (93.5)				
		Hispanic	549 (6.5)				
	Race						
		Africa/Black	364 (4.7)				
		Americas	67 (0.9)				
		Asia	850 (10.9)				
		Europe/White	6545 (83.6)				
RMET-16[a]			6338	68.0	16.0	31.3–93.8	0.56
	Country						
		Majority English/White	5056 (79.8)				
		Non-majority English/White	1281 (20.2)				
		No data	1 (<0.1)				
	Age			30.7	14.4	10–70	
	Gender						
		Female	3888 (62.1)				
		Male	2373(37.9)				
	Education						
		Middle	174 (3.0)				
		High	1504 (26.3)				
		Some college	1507 (26.4)				
		College	1383 (24.2)				
		Graduate	1147 (20.1)				
	Ethnicity						
		Non-Hispanic	5135 (92.3)				
		Hispanic	426 (7.7)				
	Race						
		Africa/Black	264 (5.1)				

(*Continued*)

https://doi.org/10.1017/S003329171800404X

Table 2. (*Continued.*)

Measure	Factor	Level	*N* (%)	*M*	S.D.	Range	α
		Americas	82 (1.6)				
		Asia	917 (17.9)				
		Europe/White	3869 (75.4)				
Multiracial Emotion Identification[b]			5213	41.1	3.5	31–46.3	0.75
	Country						
		Majority English/White	4317 (82.8)				
		Non-majority English/White	894 (17.1)				
		No data	2 (<0.1)				
	Age			29.2	14.2	10–70	
	Gender						
		Female	3153 (61.9)				
		Male	1944 (38.1)				
	Education						
		Middle	229 (4.8)				
		High	1400 (29.6)				
		Some college	1142 (24.1)				
		College	1045 (22.1)				
		Graduate	921 (19.4)				
	Ethnicity						
		Non-Hispanic	4295 (92.5)				
		Hispanic	350 (7.5)				
	Race						
		Africa/Black	203 (4.8)				
		Americas	64 (1.5)				
		Asia	607 (14.4)				
		Europe/White	3334 (79.2)				
Emotion discrimination[b]			3702	56.1	5.5	42–65	0.66
	Country						
		Majority English/White	3063 (82.7)				
		Non-majority English/White	240 (6.5)				
		No data	399 (10.8)				
	Age			28.3	13.8	10–70	
	Gender						
		Female	2458 (67.5)				
		Male	1182 (32.5)				
	Education						
		Middle	146 (4.1)				
		High	1073 (30.3)				
		Some college	1164 (32.9)				
		College	694 (19.6)				
		Graduate	466 (13.2)				
	Ethnicity						
		Non-Hispanic	3183 (94.4)				

(*Continued*)

https://doi.org/10.1017/S003329171800404X

Table 2. (*Continued.*)

Measure	Factor	Level	*N* (%)	*M*	s.d.	Range	α
		Hispanic	188 (5.6)				
	Race						
		Africa/Black	162 (5.2)				
		Americas	26 (0.8)				
		Asia	222 (7.1)				
		Europe/White	2708 (86.9)				
Digit symbol matching[b]			15 723	47.4	8.3	29–66	0.76
	Country						
		Majority English/White	11 466 (72.9)				
		Non-majority English/White	2872 (18.3)				
		No data	1386 (8.8)				
	Age			29.2	14.4	10–70	
	Gender						
		Female	7484 (48.2)				
		Male	8059 (51.8)				
	Education						
		Middle	787 (5.7)				
		High	4156 (30.0)				
		Some college	3185 (23.0)				
		College	3096 (22.3)				
		Graduate	2643 (19.0)				
	Ethnicity						
		Non-Hispanic	12 540 (92.6)				
		Hispanic	1002 (7.4)				
	Race						
		Africa/Black	653 (5.1)				
		Americas	172 (1.4)				
		Asia	2440 (19.1)				
		Europe/White	9494 (74.4)				

[a]Values represent percentage correct.
[b]Values represent number of trials correct.

Ethnicity

NIH-defined ethnicity (i.e. Hispanic or Latino ancestry) explained 8.55–26.47 times more variance in the RMET than it did in the other tasks, and only impacted performance on RMET-36, $t_{(610)} = 6.80$, $p < 0.001$, which was replicated with the RMET-16, $t_{(492)} = 6.01$, $p < 0.001$, such that non-Hispanic participants outperformed Hispanic participants (Fig. 2*b*). These differences were small-to-moderate in magnitude, RMET-36 $d = 0.33$, 95% CI 0.25–0.42, CLES = 59.28%, RMET-16 $d = 0.31$, 95% CI 0.22–0.41, CLES = 58.81%. In contrast, no difference was observed between non-Hispanic and Hispanic participants for any other task, $ts > 0.88$, $ps > 0.300$, range $d = -0.06$–0.11, CLES = 48.22–52.98%. Since effects of ethnicity are not observed with the other tasks, it suggests that the ethnicity effect may be specific to RMET performance *per se* and not social or non-social cognition more broadly.

Race

We observed an effect of race on all tasks, $Fs > 13.70$, $ps < 0.001$; however, race explained 2.31–9.89 times variance in RMET performance, RMET-36 $\eta^2 = 0.0476$, 95% CI 0.0386–0.0568, RMET-16 $\eta^2 = 0.0658$, 95% CI 0.0530–0.0787, as compared with other tasks, multiracial emotion identification $\eta^2 = 0.0207$, 95% CI 0.0126–0.0294, emotion discrimination $\eta^2 = 0.0130$, 95% CI 0.0058–0.0213, digit symbol matching $\eta^2 = 0.0067$, 95% CI 0.0040–0.0096 (Fig. 2*c*). Further, the differences between European/White *v.* non-European/non-White backgrounds differed by task. Specifically, in the RMET-36, European/White participants outperformed all other groups, ($ps < 0.001$); an effect replicated in the RMET-16 dataset ($ps < 0.001$). These differences were moderate-to-large in size, RMET-36 range $d = 0.55$–0.75, CLES = 65.02–70.32%, RMET-16 range $d = 0.56$–1.00, CLES = 65.34–76.09%, with largest differences between African/Black

https://doi.org/10.1017/S003329171800404X

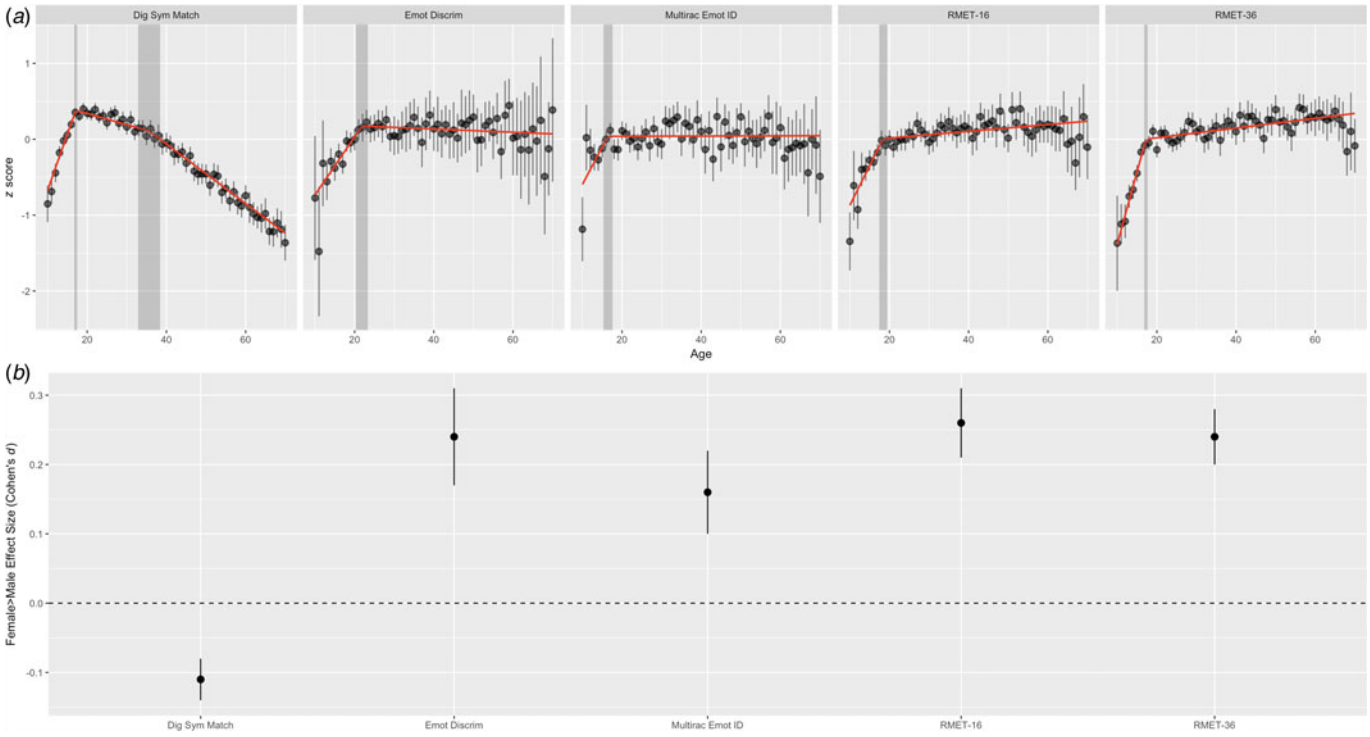

Fig. 1. (*a*) Measure performance as a function of age. The red line depicts the segmented regression slopes. The vertical gray bar depicts breakpoints and their 95% CI. Data points represent mean score as function of age with error bars denoting 95% CI. (*b*) Female>male effects sizes with 95% CI across measures. Positive values represent a female>male performance advantage; negative values represent a male>female performance advantage.

and European/White participants. In contrast, on other tasks, European/White participants less consistently outperformed non-European/non-White participants. For example, on multiracial emotion identification, European/White participants performed no differently than Black/African or Native American/Alaskan Native participants; on emotion discrimination, European/White participants performed no differently than Native American/Alaskan Native participants; on digit symbol matching, European/White participant performed no differently than Asian participants (all $ps > 0.05$). Furthermore, the effect sizes for these differences were lower for all comparisons, multiracial emotion identification range $d = 0.16–0.41$, CLES = 56.46–61.30%, emotion discrimination range $d = 0.18–0.40$, CLES = 55.06–61.23%, digit symbol matching range $d = 0.01–0.36$, CLES = 50.53–63.00%. Thus, while European/White participants tended to outperform non-European/non-White participants, this difference was only reliable and large in magnitude in the RMET compared with all other tasks.

Additional analyses

We re-ran analyses evaluating the separate effect of education, ethnicity, and race on performance, controlling for all other factors, including gender and the non-linear effect of age. All findings were unchanged.

Discussion

Findings from the current study suggest that the RMET may be unduly sensitive to several sociocultural factors over and above other social and non-social cognitive tasks. Specifically, the RMET appears to be more sensitive to demographic and sociocultural factors such that older, more highly educated individuals of non-Hispanic and White/European backgrounds are likely to outperform their younger, less educated, Hispanic, and non-White/non-European counterparts. For the other tasks, we found similar trends in that those more highly educated tended to outperform those less educated and that participants reporting European/White race tended to outperform participants reporting non-European/non-White race (although there was no effect of ethnicity on these other measures). This pattern of results whereby the influence of variables were similar in nature (although not necessarily magnitude) across tasks suggest that education and race may in fact be associated with social cognitive ability in a reliable way; a finding that may be consistent with an increasing literature documenting the deleterious effects of socioeconomic disadvantage on the brain and cognition (Zsembik and Peek, 2001; Glymour and Manly, 2008; Hackman *et al.*, 2010).

However, the effect of sociocultural variables on RMET performance is potentiated in a way that it in theory should not. Consider the finding that people with schizophrenia- or autism-spectrum disorders perform worse than healthy controls on the RMET with an effect size (Hedges' *g*) of 0.73 and 0.81, respectively (Chung *et al.*, 2014). Is there *a priori* reason to believe that on the original RMET, performance differences should be *similar* or *greater in magnitude* between participants reporting European/White *v.* African race ($d = 0.75$)? Or between participants with some college/college/graduate *v.* middle school education ($d = 0.84$, $d = 1.05$, $d = 1.03$, respectively)? Said otherwise, is it reasonable to expect these groups to perform as differently from other groups as clinical populations with marked social impairment?

In theory, some of the sociocultural-RMET performance relations could be explained in a way that has nothing to do with limitations of the task. For example, regarding the lifespan findings,

https://doi.org/10.1017/S003329171800404X

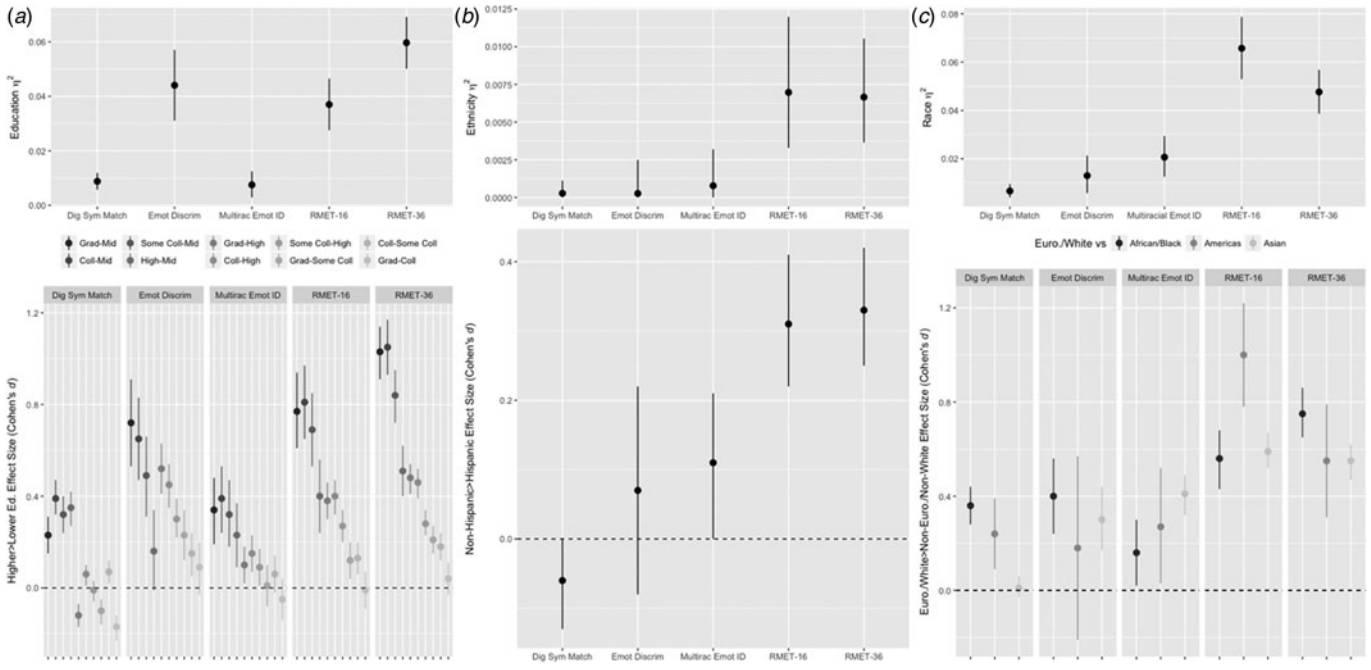

Fig. 2. (*a*) Education effects. The top panel depicts education η^2 and its 95% CI across measures. The bottom panel depicts effect size differences with 95% CI between levels of education. Positive values reflect a higher education>lower education performance advantage; negative values reflect a lower education>higher education performance advantage. Darker data points depict greater education-level differences (e.g. graduate *v.* middle school); lighter data points depict smaller education level differences (e.g. graduate *v.* college). Within each facet, education-level differences increase from right to left. (*b*) NIH-defined ethnicity effects. The top panel depicts ethnicity η^2 and its 95% CI across measures. The bottom panels depict effect size differences with 95% CI between non-Hispanic and Hispanic participants. Positive values reflect a non-Hispanic>Hispanic performance advantage; negative values reflect a Hispanic>non-Hispanic performance advantage. (*c*) Race effects. The top panel depicts race η^2 and its 95% CI across measures. The bottom panels depict effect size differences with 95% CI between European/White participants and non-European/non-White participants. Positive values reflect a European/White>non-European/non-White performance advantage; negative values reflect non-European/non-White>European/White performance advantage.

one possibility is that with age comes more social experience and expertise that would contribute to continually improving performance on the RMET over the lifespan. Something similar could be occurring with education in that more time spent in school with one's peers, navigating what might be complex social networks, may similarly positively impact social cognitive ability, explaining the large magnitude of effect that education has on RMET performance, which is consistent with other findings (Warrier *et al.*, 2018). However, if that were the case, we would expect to see similar increases in performance across the lifespan in the other social tasks, and for education to have a similar magnitude of effect on the other social tasks, which we do not. The lifespan findings with the RMET also stand in contrast to other work demonstrating a decline in social cognitive ability over the lifespan using other MSU tasks (Moran *et al.*, 2012; Moran, 2013; Klindt *et al.*, 2017). Our findings support the notion that something unique is contributing to RMET performance. It could be that age and education are serving as a proxy for verbal ability and IQ, which increases across the lifespan (Hartshorne and Germine, 2015) and has been shown to affect RMET performance (Baker *et al.*, 2014; Olderbak *et al.*, 2015); however, this is speculative given that we did not assess verbal ability or IQ directly. It is interesting, though, that performance on multiracial emotion identification, which is similar in structure (i.e. requiring emotion labeling) to the RMET, did not increase across the lifespan, and was not as related to education. This further suggests that the RMET may be particularly associated with these variables.

Regarding ethnicity and race, the RMET was the only task affected by NIH-defined ethnicity, and though White/European participants tended to outperform non-White/non-European participants across measures, these differences were less robust, and not as strong as they were with the RMET. Again, the relation between these factors and RMET could also seemingly be explained by other factors. For example, other research has revealed racial differences in aspects of mental state understanding that might affect how European/White *v.* non-European/non-White participants may approach the RMET (Masuda *et al.*, 2008; Adams *et al.*, 2010; Mason and Morris, 2010). However, if this were the case, again, we would expect to see similar performance differences among race/ethnic groups in all social tasks, which we do not. The robust racial and ethnic group differences observed on the RMET do not appear to be simply the use of non-multiracial/multiethnic stimuli either: Ethnic differences were also less reliable and robust for emotion discrimination, which similarly contained racially/ethnically homogeneous (Caucasian) stimuli. The current findings, taken with other research documenting cultural differences in performance on the RMET (Prevost *et al.*, 2014), but not other MSU tasks (Bradford *et al.*, 2018), suggest particular cultural bias with the RMET.

The RMET does appear to pick up on at least one reliable and robust demographic influence, that being gender differences. Across all social tasks, females outperformed males, which was a small, but reliable effect. This finding is consistent with other research on the RMET (Kirkland *et al.*, 2013; Baron-Cohen *et al.*, 2015; Warrier *et al.*, 2018) and other social cognitive tasks (Kret and De Gelder, 2012). This cannot be explained by general performance differences favoring females since males outperformed females on digit symbol matching.

https://doi.org/10.1017/S003329171800404X

The implications of these findings are somewhat troubling, particularly given the task's widespread use. Performance on the RMET seems to reflect aspects of social class and culture as much as it does social cognitive ability. With respect to clinical investigations, this confound can be particularly harmful. Given shared variance between social status and risk for psychiatric illness (Kendler, 1996; McLaughlin *et al.*, 2011), using the RMET, it would be difficult to tell whether performance differences are the result of psychopathology *v.* factors that covary with psychopathology, namely, socioeconomic class and culture.

How then should the field proceed? We see several ways forward. First and foremost, researchers using the RMET or tasks with similar characteristics (i.e. high reliance on vocabulary; racial and ethnic stimuli homogeneity) should be careful to consider the potentially confounding impact of sample characteristics on their findings and draw inferences with those sample characteristics in mind, particularly when comparing clinical to non-clinical groups. Second, data from the multiracial emotion identification task suggest that at least some bias related to education, race, and ethnicity can be alleviated through the use of multicultural stimuli. This notion is further supported by examining other social cognitive tasks that use multiracial/multiethnic stimuli such as the Penn Emotion Recognition Test (Kohler *et al.*, 2003; Gur and Gur, 2016), which does not appear to be affected by participant race and ethnicity (Pinkham *et al.*, 2017). Another option would be to use stimuli that are less verbally and culturally loaded, to the extent that that is possible. The Social Attribution Task-Multiple Choice task (Bell *et al.*, 2010), for example, which involves making judgments about non-verbal animated geometric objects acting with ostensible beliefs, desires, and intentions, has been validated for use across cultures (Lee *et al.*, 2018). However, other research initiatives have recommended not using the task due to poor psychometric properties (Pinkham *et al.*, 2018; although see Johannesen *et al.*, 2018). Finally, as others have recommended (Pedraza and Mungas, 2008), new assessments should be developed and validated using more diverse samples along with statistical methods that assess whether scores have similar meanings across groups (Mungas *et al.*, 2004; Siedlecki *et al.*, 2008). Moving forward, measurement initiatives such as the Social Cognition Psychometric Evaluation study (Pinkham *et al.*, 2018) will be increasingly important.

The increasing rise of large-scale cohort studies (e.g. NIH's *All of Us Research Program*), where demographic and sociocultural characteristics are likely to become increasingly crossed, means that without action toward creating sound measures, these research efforts, and others, will be profoundly undermined (Manly, 2008). Thus, we take our findings as a call to action for social cognition researchers to create measures that minimize undue effects of sociodemographic characteristics. Given ongoing initiatives at the NIMH to select and recommend social cognitive measures for widespread use in research, the time to create and test new measures is now.

Notes

[1] The *Systems for Social Processes* subgroup also recommended the Hinting Task (Corcoran *et al.*, 1995) as the best option for logical/physical perspective taking.

[2] For example, correct response options for female faces include words with sexual connotations such as *desire*, *flirtatious*, and *fantasizing*; correct response options for male faces include words with aggression or power connotations such as *insisting*, *accusing*, *defiant*, *hostile*, and *serious*. Additionally, as determined with Linguistic Inquiry and Word Count (LIWC) software (Pennebaker *et al.*, 2015), correct response options for female faces include more positive emotion words (41.8%) than for male faces (10.5%), and correct response options for male faces include more negative emotion words (36.8%) than for female faces (17.7%).

[3] We chose digit symbol matching as the non-social task because of its non-specific sensitivity to a variety of demographic factors, for example, age (Hartshorne and Germine, 2015), intelligence and cognition (Salthouse, 1996), neurological health (Longstreth *et al.*, 2005), mental health (Dickinson *et al.*, 2007), and mortality (Fried, 1998). Said otherwise, performance on this task appears to have widespread *robust* and *meaningful* correlates. Thus, associations between digit symbol matching performance and demographic factors may serve as an informal baseline for possible relations between performance on the other measures and demographic factors.

[4] We used the Corpus of Contemporary American English (Davies, 2010) to calculate the word frequency of the response options for the RMET, the emotion identification measure (*happy*, *sad*, *angry*, *fearful*), and the emotion discrimination measure (*same*, *different*). The response options for RMET were massively less frequent compared with the response options for the emotion identification task, $d = 2.1$, 95% CI (1.02–3.13), common language effect size (CLES) = 92.9% and emotion discrimination task, $d = 22.3$, 95% CI (18.8–25.8), CLES = 100%. Said otherwise, there is a 93% chance that a randomly selected response option from the emotion identification task will be more frequent than a randomly selected word from the RMET; and there is a 100% chance that a randomly selected response option from the emotion discrimination task will be more frequent than a randomly selected response option from the RMET.

[5] Negative effect size indicates that a group with less education outperformed a group with more education.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S003329171800404X.

Author ORCIDs. David Dodell-Feder, 0000-0003-0678-1728

Acknowledgements. We thank Erin Dunn and Andrew Smith for advice on data analysis, and the TestMyBrain.org volunteers for their participation.

Financial support. This project was supported by the Stanley Center for Psychiatric Research at the Broad Institute of MIT and Harvard as well as through NIH National Institute of Mental Health contract HHSN271201700776P awarded to L Germine.

Conflict of interest. None.

Ethical standards. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

References

Adams RB, Rule NO, Franklin RG, Wang E, Stevenson MT, Yoshikawa S, Nomura M, Sato W, Kveraga K and Ambady N (2010) Cross-cultural reading the mind in the eyes: an fMRI investigation. *Journal of Cognitive Neuroscience* **22**, 97–108.

Anagnostou E, Soorya L, Chaplin W, Bartz J, Halpern D, Wasserman S, Wang AT, Pepa L, Tanel N, Kushki A and Hollander E (2012) Intranasal oxytocin versus placebo in the treatment of adults with autism spectrum disorders: a randomized controlled trial. *Molecular Autism* **3**, 16.

Baker CA, Peterson E, Pulos S and Kirkland RA (2014) Eyes and IQ: a meta-analysis of the relationship between intelligence and 'Reading the Mind in the Eyes'. *Intelligence* **44**, 78–92.

Baron-Cohen S, Wheelwright S, Hill J, Raste Y and Plumb I (2001*a*) The 'Reading the Mind in the Eyes' test revised version: a study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry* **42**, 241–251.

Baron-Cohen S, Wheelwright S, Spong A, Scahill V and Lawson J (2001*b*) Are intuitive physics and intuitive psychology independent? A test with children with Asperger Syndrome. *Journal of Developmental and Learning Disorders* **5**, 47–78.

Baron-Cohen S, Bowen DC, Holt RJ, Allison C, Auyeung B, Lombardo MV, Smith P and Lai M-C (2015) The 'Reading the Mind in the Eyes' test: complete absence of typical sex difference in ~400 men and women with autism. Ed. H Yamasue *PLoS ONE* **10**, e0136521.

Bell MD, Fiszdon JM, Greig TC and Wexler BE (2010) Social attribution test – multiple choice (SAT-MC) in schizophrenia: comparison with community sample and relationship to neurocognitive, social cognitive and symptom measures. *Schizophrenia Research* **122**, 164–171.

Black JE (2018) An IRT analysis of the reading the mind in the eyes test. *Journal of Personality Assessment*, 1–9. doi: 10.1080/00223891.2018.1447946.

Blatt B, LeLacheur SF, Galinsky AD, Simmens SJ and Greenberg L (2010) Does perspective-taking increase patient satisfaction in medical encounters? *Academic Medicine* **85**, 1445–1452.

Bora E and Pantelis C (2013) Theory of mind impairments in first-episode psychosis, individuals at ultra-high risk for psychosis and in first-degree relatives of schizophrenia: systematic review and meta-analysis. *Schizophrenia Research* **144**, 31–36.

Bora E, Yucel M and Pantelis C (2009) Theory of mind impairment in schizophrenia: meta-analysis. *Schizophrenia Research* **109**, 1–9.

Bora E, Bartholomeusz C and Pantelis C (2016) Meta-analysis of Theory of Mind (ToM) impairment in bipolar disorder. *Psychological Medicine* **46**, 253–264.

Boyd D, Hargittai E, Schultz J and Palfrey J (2011) Why parents help their children lie to Facebook about age: unintended consequences of the 'Children's Online Privacy Protection Act'. *First Monday* **16**. Available at https://journals.uic.edu/ojs/index.php/fm/article/view/3850/3075.

Bradford EEF, Jentzsch I, Gomez J-C, Chen Y, Zhang D and Su Y (2018) Cross-cultural differences in adult theory of mind abilities: a comparison of native-English speakers and native-Chinese speakers on the self/other differentiation task. *Quarterly Journal of Experimental Psychology*, **71**, 2665–2676.

Chung YS, Barch D and Strube M (2014) A meta-analysis of mentalizing impairments in adults with schizophrenia and autism spectrum disorder. *Schizophrenia Bulletin* **40**, 602–616.

Corcoran R, Mercer G and Frith CD (1995) Schizophrenia, symptomatology and social inference: investigating 'theory of mind' in people with schizophrenia. *Schizophrenia Research* **17**, 5–13.

Davies M (2010) The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* **25**, 447–464.

Dickinson D, Ramsey ME and Gold JM (2007) Overlooking the obvious: a meta-analytic comparison of digit symbol coding tasks and other cognitive measures in schizophrenia. *Archives of General Psychiatry* **64**, 532.

Dodell-Feder D and Tamir DI (2018) Fiction reading has a small positive impact on social cognition: a meta-analysis. *Journal of Experimental Psychology: General* **147**, 1713–1727.

Dodell-Feder D, Lincoln SH, Coulson JP and Hooker CI (2013) Using fiction to assess mental state understanding: a New task for assessing theory of mind in adults. Ed. L Young *PLoS ONE* **8**, e81279.

Dodell-Feder D, DeLisi LE and Hooker CI (2014*a*) Neural disruption to theory of mind predicts daily social functioning in individuals at familial high-risk for schizophrenia. *Social Cognitive and Affective Neuroscience* **9**, 1914–1925.

Dodell-Feder D, Tully LM, Lincoln SH and Hooker CI (2014*b*) The neural basis of theory of mind and its relationship to social functioning and social anhedonia in individuals with schizophrenia. *NeuroImage: Clinical* **4**, 154–163.

Dodell-Feder D, Felix S, Yung MG and Hooker CI (2016) Theory-of-mind-related neural activity for one's romantic partner predicts partner well-being. *Social Cognitive and Affective Neuroscience* **11**, 593–603.

Fink E, Begeer S, Peterson CC, Slaughter V and de Rosnay M (2015) Friendlessness and theory of mind: a prospective longitudinal study. *British Journal of Developmental Psychology* **33**, 1–17.

Fried LP (1998) Risk factors for 5-year mortality in older AdultsThe cardiovascular health study. *JAMA* **279**, 585.

Garrido L, Furl N, Draganski B, Weiskopf N, Stevens J, Tan GC-Y, Driver J, Dolan RJ and Duchaine B (2009) Voxel-based morphometry reveals reduced grey matter volume in the temporal cortex of developmental prosopagnosics. *Brain* **132**, 3443–3455.

Germine L and Hooker CI (2011) Face emotion recognition is related to individual differences in psychosis-proneness. *Psychological Medicine* **41**, 937–947.

Germine L, Garrido L, Bruce L and Hooker C (2011) Social anhedonia is associated with neural abnormalities during face emotion processing. *NeuroImage* **58**, 935–945.

Germine L, Nakayama K, Duchaine BC, Chabris CF, Chatterjee G and Wilmer JB (2012) Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review* **19**, 847–857.

Glymour MM and Manly JJ (2008) Lifecourse social conditions and racial and ethnic patterns of cognitive aging. *Neuropsychology Review* **18**, 223–254.

Goldstein NJ, Vezich IS and Shapiro JR (2014) Perceived perspective taking: when others walk in our shoes. *Journal of Personality and Social Psychology* **106**, 941–960.

Gur RC and Gur RE (2016) Social cognition as an RDoC domain. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **171**, 132–141.

Hackman DA, Farah MJ and Meaney MJ (2010) Socioeconomic status and the brain: mechanistic insights from human and animal research. *Nature Reviews Neuroscience* **11**, 651–659.

Hartshorne JK and Germine L (2015) When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science* **26**, 433–443.

Holt-Lunstad J, Smith TB, Baker M, Harris T and Stephenson D (2015) Loneliness and social isolation as risk factors for mortality: a meta-analytic review. *Perspectives on Psychological Science* **10**, 227–237.

Imuta K, Henry JD, Slaughter V, Selcuk B and Ruffman T (2016) Theory of mind and prosocial behavior in childhood: a meta-analytic review. *Developmental Psychology* **52**, 1192–1205.

Johannesen JK, Fiszdon JM, Weinstein A, Ciosek D and Bell MD (2018) The Social Attribution Task – Multiple Choice (SAT-MC): psychometric comparison with social cognitive measures for schizophrenia research. *Psychiatry Research* **262**, 154–161.

Johnston L, Miles L and McKinlay A (2008) A critical review of the eyes test as a measure of social-cognitive impairment. *Australian Journal of Psychology* **60**, 135–141.

Kendler KS (1996) Lifetime prevalence, demographic risk factors, and diagnostic validity of nonaffective psychosis as assessed in a US community sample: The National Comorbidity Survey. *Archives of General Psychiatry* **53**, 1022.

Kim HS, Shin NY, Jang JH, Kim E, Shim G, Park HY, Hong KS and Kwon JS (2011) Social cognition and neurocognition as predictors of conversion to psychosis in individuals at ultra-high risk. *Schizophrenia Research* **130**, 170–175.

Kirkland RA, Peterson E, Baker CA, Miller S and Pulos S (2013) Meta-analysis reveals adult female superiority in 'Reading the Mind in the Eyes Test'. *North American Journal of Psychology* **15**, 121–146.

Klindt D, Devaine M and Daunizeau J (2017) Does the way we read others' mind change over the lifespan? Insights from a massive web poll of cognitive skills from childhood to late adulthood. *Cortex* **86**, 205–215.

Kohler CG, Turner TH, Bilker WB, Brensinger CM, Siegel SJ, Kanes SJ, Gur RE and Gur RC (2003) Facial emotion recognition in schizophrenia: intensity effects and error pattern. *American Journal of Psychiatry* **160**, 1768–1774.

Kret ME and De Gelder B (2012) A review on sex differences in processing emotional signals. *Neuropsychologia* **50**, 1211–1221.

Lee H-S, Corbera S, Poltorak A, Park K, Assaf M, Bell MD, Wexler BE, Cho Y-I, Jung S, Brocke S and Choi K-H (2018) Measuring theory of mind in schizophrenia research: cross-cultural validation. *Schizophrenia Research*, **201**, 187–195.

Longstreth WT, Arnold AM, Beauchamp NJ, Manolio TA, Lefkowitz D, Jungreis C, Hirsch CH, O'Leary DH and Furberg CD (2005) Incidence, manifestations, and predictors of worsening white matter on serial cranial magnetic resonance imaging in the elderly: the cardiovascular health study. *Stroke* **36**, 56–61.

Manly JJ (2008) Critical issues in cultural neuropsychology: profit from diversity. *Neuropsychology Review* **18**, 179–183.

Mason MF and Morris MW (2010) Culture, attribution and automaticity: a social cognitive neuroscience view. *Social Cognitive and Affective Neuroscience* **5**, 292–306.

Masuda T, Ellsworth PC, Mesquita B, Leu J, Tanida S and Van de Veerdonk E (2008) Placing the face in context: cultural differences in the perception of facial emotion. *Journal of Personality and Social Psychology* **94**, 365–381.

McDonald S, Flanagan S, Rollins J and Kinch J (2003) TASIT: a new clinical tool for assessing social perception after traumatic brain injury. *The Journal of Head Trauma Rehabilitation* **18**, 219–238.

McLaughlin KA, Breslau J, Green JG, Lakoma MD, Sampson NA, Zaslavsky AM and Kessler RC (2011) Childhood socio-economic status and the onset, persistence, and severity of DSM-IV mental disorders in a US national sample. *Social Science & Medicine* **73**, 1088–1096.

Molenberghs P, Johnson H, Henry JD and Mattingley JB (2016) Understanding the minds of others: a neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews* **65**, 276–291.

Moran JM (2013) Lifespan development: the effects of typical aging on theory of mind. *Behavioural Brain Research* **237**, 32–40.

Moran JM, Jolly E and Mitchell JP (2012) Social-cognitive deficits in normal aging. *Journal of Neuroscience* **32**, 5553–5561.

Muggeo VMR (2003) Estimating regression models with unknown break-points. *Statistics in Medicine* **22**, 3055–3071.

Muggeo VMR (2008) Segmented: an R package to fit regression models with broken-line relationships. *R News* **8**, 20–25.

Mungas D, Reed BR, Crane PK, Haan MN and González H (2004) Spanish and English Neuropsychological Assessment Scales (SENAS): further development and psychometric characteristics. *Psychological Assessment* **16**, 347–359.

National Advisory Mental Health Council Workgroup on Tasks and Measures for RDoC (2016) *Behavioral Assessment Methods for RDoC Constructs: A Report by the National Advisory Mental Health Council Workgroup on Tasks and Measures for Research Domain Criteria* (*RDoC*).

Nisbett RE (2004) *The Geography of Thought: How Asians and Westerners Think Differently … and Why*. New York: Nachdr. Free Press.

Olderbak S, Wilhelm O, Olaru G, Geiger M, Brenneman MW and Roberts RD (2015) A psychometric analysis of the reading the mind in the eyes test: toward a brief form for research and applied settings. *Frontiers in Psychology* **6**, 1503.

Pedraza O and Mungas D (2008) Measurement in cross-cultural neuropsychology. *Neuropsychology Review* **18**, 184–193.

Pennebaker JW, Boyd RL, Joran K and Blackburn K (2015) *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Pinkham AE, Penn DL, Green MF, Buck B, Healey K and Harvey PD (2014) The social cognition psychometric evaluation study: results of the expert survey and RAND panel. *Schizophrenia Bulletin* **40**, 813–823.

Pinkham AE, Harvey PD and Penn DL (2018) Social cognition psychometric evaluation: results of the final validation study. *Schizophrenia Bulletin* **44**, 737–748.

Pinkham AE, Kelsven S, Kouros C, Harvey PD and Penn DL (2017) The effect of age, race, and sex on social cognitive performance in individuals with schizophrenia. *The Journal of Nervous and Mental Disease* **205**, 346–352.

Pitcher D, Garrido L, Walsh V and Duchaine BC (2008) Transcranial magnetic stimulation disrupts the perception and embodiment of facial expressions. *Journal of Neuroscience* **28**, 8929–8933.

Prevost M, Carrier M-E, Chowne G, Zelkowitz P, Joseph L and Gold I (2014) The reading the mind in the eyes test: validation of a French version and exploration of cultural variations in a multi-ethnic city. *Cognitive Neuropsychiatry* **19**, 189–204.

R Core Team (2017) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Salthouse TA (1996) The processing-speed theory of adult age differences in cognition. *Psychological Review* **103**, 403–428.

Salthouse TA (2004) What and when of cognitive aging. *Current Directions in Psychological Science* **13**, 140–144.

Siedlecki KL, Tucker-Drob EM, Oishi S and Salthouse TA (2008) Life satisfaction across adulthood: different determinants at different ages? *The Journal of Positive Psychology* **3**, 153–164.

Slaughter V, Imuta K, Peterson CC and Henry JD (2015) Meta-analysis of theory of mind and peer popularity in the preschool and early school years. *Child Development* **86**, 1159–1174.

Todd AR and Galinsky AD (2014) Perspective-taking as a strategy for improving intergroup relations: evidence, mechanisms, and qualifications: perspective-taking and intergroup relations. *Social and Personality Psychology Compass* **8**, 374–387.

Todd AR, Bodenhausen GV, Richeson JA and Galinsky AD (2011) Perspective taking combats automatic expressions of racial bias. *Journal of Personality and Social Psychology* **100**, 1027–1042.

Walker ER, McGee RE and Druss BG (2015) Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. *JAMA Psychiatry* **72**, 334.

Warrier V, Grasby KL, Uzefovsky F, Toro R, Smith P, Chakrabarti B, Khadake J, Mawbey-Adamson E, Litterman N, Hottenga J-J, Lubke G, Boomsma DI, Martin NG, Hatemi PK, Medland SE, Hinds DA, Bourgeron T and Baron-Cohen S (2018) Genome-wide meta-analysis of cognitive empathy: heritability, and correlates with sex, neuropsychiatric conditions and cognition. *Molecular Psychiatry* **23**, 1402–1409.

Zsembik BA and Peek MK (2001) Race differences in cognitive functioning among older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* **56**, S266–S274.