Inclusive Neuropsychological and Genomic Analysis of Social Behaviors in Autism Spectrum Disorder and Schizophrenia

Heesu (Ally) Kim

Submitted in Partial Fulfillment of the Prerequisite for Honors in the Department of Neuroscience under the advisement of Jeremy B. Wilmer and Laura Germine

May 2021

© 2021 Heesu (Ally) Kim
# Table of Contents

Abstract  
Acknowledgements  
Introduction  
Chapter 1: Multiracial Reading the Mind in the Eyes Test (MRMET)  
  Background  
  Materials and Methods  
  Results  
  Discussion  
  Conclusion  
Chapter 2: Primer for Behavioral Genetics and Genotype-Phenotype Analysis  
  Introduction  
  Biological Spectrum in ASD and SCZ  
  Genome-Wide Association Studies  
  ASD and SCZ Polygenic Risk Score  
  Limitations of Current Approaches  
  Summary  
Chapter 3: Genetic Correlates of Social Phenotypes in ASD and SCZ  
  Background  
  Methods  
  Results  
  Discussion and Future Direction  
  Conclusion  
Literature Cited  
Author Contributions  
Supplemental Information  
Appendix
Abstract

Science that has shaped our current knowledge of psychiatric conditions has often relied on non-diverse, non-inclusive methods and samples. Typically, studies of autism spectrum disorder (ASD) and schizophrenia (SCZ), including both phenotypic investigations — which use behavioral tests — and genotypic investigations — which employ population-based genetic data — have relied on tests and/or samples that resemble only the Caucasian majority. In order to investigate how diversity could be more firmly baked into the fabric of psychiatric research, I investigated a widely-used phenotypic measure for social cognition ASD and SCZ, called the Reading the Mind in the Eyes Test (RMET). I show that an improved successor, the Multiracial Reading the Mind in the Eyes test (MRMET), fares equally well or better than the Caucasian version on multiple fronts. Second, I surveyed the process of computing non-categorical genomic risk scores, called the polygenic risk score (PRS), and asked how sample selection limitations typical of genomic research may lead to decreased generalizability of the PRS. Third, I assessed how the limitations in both phenotypic and genetic research may come together to affect scientific conclusions in joint genotype-phenotype research. Finally, I assess how the integration of typical genotypic (PRS) and phenotypic (behavioral factor) information yield informative yet distinct results in a European and a non-European sample. Together, these studies point the way towards improved diversity and systematic inclusivity in modern scientific approaches that aim to define diseases of human behavior.
Acknowledgements

First and foremost, I would like to thank my thesis committee for their relentless support and faith in me and throughout this process. I would like to particularly thank Professor Jeremy B. Wilmer and Dr. Laura Germine for the three years of close research supervision, instruction, and inspiration, as I know I could not have gotten to where I am—both as a scientist and a person—without them. I would like to thank Dr. Deborah Bauer for providing me—as a budding scientist—the resources and the community to pursue my first research project three years ago. This formative experience has now led me to be where I am today. Finally, I would also like to thank Dr. Elise Robinson at Broad Institute for enabling me to pursue this rich topic through the resources—including data—from her laboratory, and ultimately enabling me to continue this important project in the Robinson Laboratory upon graduation.

I would like to thank my entire Thesis Committee for advice, expertise, and support, which further includes Dr. Cassandra Pattanayak, whose statistical insights were invaluable, and Dr. Phil H. Lee whose teachings in the field of genomic medicine was foundational for this project.

I would like to thank Dr. Peggy Levitt, Dr. Claudia Malone, and the Office of the Provost for funding my thesis research as well as research in years prior in the laboratory I am currently working in. I would also like to thank the Goldwater Foundation for partially supporting and funding my last year of college as I pursued this project.

I would like to thank my parents, who are immigrants, for funding my education and allowing me to be here. I would also like to thank my high school biology teacher, Mrs. Lenz, for fostering a love for biology and science in me that is sure to grow me into an excellent physician-scientist in the future.

Finally, I would like to thank my friends for inspiring me and encouraging me in my greatest times of need.
Introduction

This thesis takes as its central focus that any flaws embedded within the science of describing people have real-life consequences. Notably, scientific findings can be only as good as the data and the research methods themselves, which thereby necessitate the employment of good tests and good experiments. As we, human scientists, develop tools, methodologies, and measurements to characterize and understand people at both the behavioral and biological level, we must bear in mind the critical role that such tools may play in placing, characterizing, categorizing, and empowering people in the broader society. To that end, this paper consists of three parts.

First, I approach the Reading the Mind in the Eyes Test (RMET)—a seminal emotion identification task for autism research—in light of what makes a neurocognitive test “good.” In this section, I focus specifically on the potential for greater diversity in psychopathology-related measurement: whether multiracial cognitive assessments can fare as well as, if not better than, racially homogenous cognitive assessments. To answer this question, I employ data from over 27,000 TestMyBrain.org participants who have taken the RMET or the Multiracial RMET (MRMET) to compare both tests across race, age, education, and performance on other relevant cognitive tasks. Second, I survey one computational method that holds promise in understanding genotypic-phenotypic relationships—the polygenic risk score—for conditions like ASD and schizophrenia. I put into perspective the limitations of current genomic research and generalizability of the applications of genomics, such as using polygenic risk scores on non-European and non-Asian populations. Third, I investigate—from a dataset compiled across more than 1,300 individuals from the Simon Simplex Collection—how genomic risk for ASD and schizophrenia are involved in two facets of social ability. Together, these chapters emphasize
the importance of measurement tools that focus on not only reliability, validity, and effect size estimation, but also diversity and inclusion for use in neuropsychological or genomic research.
Chapter 1: Multiracial Reading the Mind in the Eyes Test (MRMET)

Background

Psychological research is only as accurate as the behavioral assessments themselves. As one example, to understand genetic correlates of social ability in autism spectrum disorder (ASD) or schizophrenia (SCZ), it is important to develop robust and precise behavioral tasks that accurately measure social ability. Despite the need for good, generalizable behavioral tests, many seminal cognitive tests that involve facial images still mostly revolve around Caucasian faces. As just one illustrative example, a recent paper “removed 247 non-Caucasian faces (to avoid well-known other-race effects in face perception […] )” (Sutherland et al., 2020). This example is instructive in part because it explicitly states a common, but often unstated, driving force in this domain: the assumption diversity of stimuli might compromise experimental control, and therefore reduce the reliability or validity of measurement. Our work here will specifically falsify this assumption in an influential domain of social cognitive measurement, and we will present this example as a case-in-point of the importance of questioning that assumption in other domains as well.

Interestingly, even attempts to introduce diversity in neuropsychological stimuli have often resulted in continued homogenization of the races of presented faces (that is, the “diverse” tests still often rely only on one particular race). Examples of this phenomenon include the Asian Cambridge Face Memory Test (McKone et al., 2017); Black (Handley et al., 2019) or Asian (Adams et al., 2010) Reading the Mind in the Eyes Tests; and the Japanese and Caucasian Brief Affect Recognition Test (JACBART; Matsumoto et al., 2000).
A few questions remain: First, can multiracial cognitive assessments that utilize a diverse set of stimuli maintain psychometric qualities that are as good as, if not better than, Eurocentric measures? Second, Does data support these instincts toward experimental control that have led us to create racially homogenous face tests across various races? Lastly, does the homogenization of stimuli in fact yield more accurate information than heterogeneous stimuli? On one hand, a stimulus set of varying skin colors, genders, and identity expressions introduces potential confounds in resulting data. Such potential confounds include other-race effects on reaction time (Dahl et al., 2014; Ge et al., 2009), facial recognition ability (Ng & Lindsay, 1994), and effect of implicit biases (Hugenberg & Bodenhausen, 2003), which each have been demonstrated time and time again (Alexandre et al., 2018). On the other hand, introducing diversity in cognitive assessments may be a more ethological alternative to highly curated and unnatural datasets such that a diverse stimulus set is more likely to predict real-life behavior (Chen et al., 2020) while remaining inclusive of cultural diversity (Dodell-Feder et al., 2020).

Ultimately, there is a lack of consensus on whether diverse stimuli perform better or worse than homogenous ones for the purposes of non-race-specific scientific inquiry.

To answer the question of how well a multiracial face test performs compared to a racially homogeneous one, we study the Reading the Mind in the Eyes Test (RMET) as a model example. The RMET is an emotion identification task that is widely used in neuropsychiatric literature and clinical practice to estimate impairments in social cognitive ability (Baron-Cohen et al., 2001; Vallente et al., 2013). The RMET is increasingly influencing important projects in varied fields. For instance, the RMET has thus far been cited at least 5,400 times according to Google Scholar, with increasing momentum each year demonstrated by its increasing number of new citations (Baron-Cohen et al., 2001, Figure 1.1).
Figure 1.1: Citations per year of the Reading the Mind in the Eyes Test. Since its creation in 2001, the Reading the Mind in the Eyes Test is increasingly cited each year. Citations estimates are sourced from Google Scholar metrics (scholar.google.com). Years 2001 and 2020 are omitted due to incomplete data for the corresponding one-year periods.

The RMET has now been adapted for better understanding psychopathology (Rutherford et al., 2002; Baron-Cohen et al., 2015; Holt et al., 2014; Kettle et al., 2008; Pickup, 2008; De Achaval, 2010), development and aging (Moor et al., 2012; Castelli et al., 2010; Hartshorne & Germine, 2015), and genes and environment (Rodrigues et al., 2009; Germine et al., 2015; Simon et al., 2019). RMET’s reach goes beyond purely behavioral investigations, influencing molecular psychiatry (Guastella et al., 2013), neuroscience (Kynast et al., 2020), and even genomics (Warrier et al., 2013). Prominently, the RMET is one of the two tests recommended for use in future research by the National Institute of Mental Health (NIMH) as a part of the Research Domain Criteria (RDoC), a well-known initiative that is driving the future of mental health research (Sanislow et al., 2010). Given the prominence of RMET as a measurement for autism-related social behaviors, it is clear that RMET has thus shaped the field of social behavior in autism and psychopathology.

However, the RMET is not without its critical limitations. First, the task solely relies on pictures of white, European faces to assess emotion recognition ability, presenting a critical source of systematic bias. Beyond its limited focus on white faces, the RMET also has answer
choices that are highly stereotyped (for example, the words “fantasizing” and “desire” are the correct answers for multiple pictures of women) or difficult to understand (“aghast”). Third, the correct answers for all questions were decided upon by a small group of test developers, not by the person experiencing the emotion. In this section, we further explore the limitations of the RMET in detail.

**Limitation I: Racial homogeneity.** The benefits of diverse representation in psychology have been well-established as it enables rapport between the subjects and the researchers while yielding results that are ethologically valid (Shriver et al., 2007; Golby et al., 2001; Reynolds & Suzuki, 2012; Pedraza & Mungas, 2008; Olson & Jacobson, 2015). RMET, however, employs a homogeneous set of stimuli: distinctively white, European faces. Its success as a measure has led to it being translated into multiple languages and used in cross-cultural settings and with a diverse array of populations (Sanvicente-Vieira et al., 2014). Particularly in light of this increased breadth of use, important questions arise about its lack of diverse stimuli representative of the range of potential test takers. These restricted stimuli plausibly limit the generalizability of the RMET, raising questions about its validity for such global or cross-cultural use. Moreover, they risk alienating or distressing participants who are not reflected in the stimuli, a concern expressed by participants who have participated in experiments utilizing the measure (**Table 1.1**).

**Limitation II: Gender stereotyping.** The RMET makes heavy use of gender stereotyped images and word choices, reflective of the magazine clippings and other print media from which the stimuli were drawn. For example, women’s eyes have heavy makeup use, and the word choices for female—but not male—images commonly include words such as
“fantasizing” and “flirtatious.” Of the 17 female images, 11 included commonly
gender-stereotyped words as the correct answers, suggesting the daunting possibility that
adherence to gender stereotypes may actually lead to better scores on the RMET
(Supplemental Table 1.1, Table 1.1). In addition to constituting potential threats to validity,
gender stereotypes portrayed by the RMET have the potential to alienate, distress, or bias a
subset of participants as denoted in Table 1.2.

Table 1.1: Correct answer choices in the RMET for male and
female-presenting faces across the stimulus set. Words for females
include “desire,” “fantasizing,” and “interested,” which conform to gender stereotypes.

<table>
<thead>
<tr>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>accusing</td>
<td>cautious</td>
</tr>
<tr>
<td>anticipating</td>
<td>confident</td>
</tr>
<tr>
<td>cautious</td>
<td>contemplative</td>
</tr>
<tr>
<td>concerned</td>
<td>decisive</td>
</tr>
<tr>
<td>defiant</td>
<td>desire</td>
</tr>
<tr>
<td>despondent</td>
<td>distrustful</td>
</tr>
<tr>
<td>friendly</td>
<td>doubtful</td>
</tr>
<tr>
<td>hostile</td>
<td>fantasizing (2)</td>
</tr>
<tr>
<td>insisting</td>
<td>flirtatious</td>
</tr>
<tr>
<td>panicked</td>
<td>interested (2)</td>
</tr>
<tr>
<td>pensive</td>
<td>nervous</td>
</tr>
<tr>
<td>playful</td>
<td>preoccupied (2)</td>
</tr>
<tr>
<td>regretful</td>
<td>reflective</td>
</tr>
<tr>
<td>sceptical</td>
<td>tentative</td>
</tr>
<tr>
<td>serious</td>
<td></td>
</tr>
<tr>
<td>suspicious</td>
<td></td>
</tr>
<tr>
<td>thoughtful</td>
<td></td>
</tr>
<tr>
<td>uneasy</td>
<td></td>
</tr>
<tr>
<td>upset</td>
<td></td>
</tr>
<tr>
<td>worried</td>
<td></td>
</tr>
</tbody>
</table>
Table 1.2: Reading the Mind in the Eyes Test participants mention racial and gender biases present in the measure.

<table>
<thead>
<tr>
<th>Sample Participant Feedback on the RMET</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Racial homogeneity</strong></td>
</tr>
<tr>
<td>&quot;The mind in the eyes was a bit challenging, and am not sure if this could be because it was all white faces, and I could probably not relate easily with the expressions!&quot;</td>
</tr>
<tr>
<td>&quot;This patient shared feedback about diversity concerns. He had asked during enrollment and his initial check-in about the importance of diversity in the study (meaning, does the study value diversity?) and was assured both times that this was the case. … There was not a single image of a person of color, which can feel discouraging/unwelcoming and he is now feeling ambivalent about continuing in the study.&quot;</td>
</tr>
<tr>
<td><strong>Gender stereotyping</strong></td>
</tr>
<tr>
<td>&quot;[It] was a good test but I thought that the emotions through eyes text was sexist as any emotion to do with desire or fear I thought, was a woman and any emotion which was to do with power or kindness was from a man.&quot;</td>
</tr>
<tr>
<td>&quot;As I previously stated, your eyes/emotion test is total sexist [<strong><strong>]. Not all women wear a [</strong></strong>]-ton of makeup and like to be used for male fantasy emotions like &quot;desire&quot;, etc. Why don't you hire some female researchers. Shame on you and your male sexist bias. Shame.”</td>
</tr>
<tr>
<td>&quot;I thought the photos with the emotions were rather gender biased i.e reflected media stereotypes&quot;</td>
</tr>
</tbody>
</table>

Comments were collected as feedback from TestMyBrain.org upon completion of the RMET. Participants note sexism and nondiversity as primary concerns for the measure. **** denote censored expletives.

**Limitation III: Consensus scoring.** A potentially important distinction can be made between scoring based on the similarity of one's responses to those of a reference group—also called consensus scoring—versus scoring based on an objective ground truth. This distinction may be all the more important in a domain like face processing where the life experiences of a test taker may differ markedly from those of a reference group. For the RMET, stimuli were
taken from popular print media, target words were selected by the test developers themselves, and the correctness of the target words relative to the foil words was initially determined via consensus among a small set of eight judges recruited by the authors (Baron-Cohen et al., 2001). Each of these three steps in the development of the RMET raises important questions about the extent to which the RMET taps an underlying ground truth in what was going through the mind of the person making the expression.

Multiracial Reading the Mind in the Eyes. Given these limitations, we therefore consider the development of a multiracial, non-gendered, ground-truth linked RMET a potentially valuable contribution. Here, we present such a revised version of the classic RMET that has been developed and validated through TestMyBrain.org (TMB), a digital research platform. We call this test the TestMyBrain Multiracial Reading the Mind in the Eyes Test (TMB-MRMET, or just MRMET, Figure 1.2). Underlying this work is the deeper question of whether diverse stimuli can be woven into a successful neurocognitive measure in a way that does not compromise its effectiveness as a measure. We therefore implement a head-to-head comparison of MRMET and RMET, asking whether MRMET demonstrates comparable levels of reliability and validity to RMET. We find, in a series of large studies utilizing diverse participant samples, that MRMET performs as well as or better than RMET in all domains. Moreover, MRMET captures variation that is nearly indistinguishable from the RMET. We conclude that MRMET not only removes three key limitations of RMET in terms of its stimuli but does so without compromising psychometrically, therefore representing a strong alternative to RMET with wide potential utility. In sum, I showcase the necessity of reevaluating common behavioral assessments and the assumptions underlying such assessments—a cause that is particularly
important in the process of democratizing science and “scaling up” cognitive assessments to a larger population for integration with, as one example, genomic databases.

Figure 1.2: Comparison of the Reading the Mind in the Eyes Test (RMET, top) (Baron-Cohen et al., 2001) and the Multiracial Reading the Mind in the Eyes Test Test (MRMET, bottom). A normalized density plot (right) of the score distribution for each task is provided in units of proportion correct.
Materials and Methods

**Stimuli.** Our goal was to create a multiracial, age-diverse, and less stereotypic set of stimuli with words at an accessible reading level. We generated a set of 37 items based on videos and images of racially and age-diverse actors expressing different emotions, taken as part of the Act Out for Brain Health project. We recruited professional actors from across a range of ages and races from the Boston theatre community. The actors were given mental state words to depict via their facial expressions and were recorded while holding a sheet of paper with the word they were depicting. Individual images for each word were extracted from the video recordings, cropped to include eyes and eyebrows, and compiled into a database. Post-face bank creation, we worked on paring down the literacy level of the original set of words, and removed commonly gendered words such as “fantasizing” and “flirtatious.” American literacy rates were not accounted for when the original test creators assigned words to images; roughly a third of Americans read at or below a 5th grade reading level (Mamedova & Pawlowski 2019), while the original test sits at a college reading level, as indicated by Flesch–Kincaid Grade Level (see Baron-Cohen et al., 2001 for a full list of words). In an effort to increase accessibility, we removed words beyond a fifth grade reading level, as indicated by standardized testing parameters for fifth graders in 2018.

**Iterative Task Development.** For validation, we used a set of 16 core items from the original test, identified based on items loading highest on the first principal component derived from principal component analysis, and administered a modified version of the test with these 16 core items and a random set of 37 out of 109 new items. These methods allowed us to identify those items with the highest correlations with the original test items to generate a new test that avoids many of the controversial characteristics of the original test, such as high
correlations with vocabulary and gendered language. Over 6,000 participants completed the test, with an average of 1,300 participants per item. The final set of stimuli were 37 items selected based on convergence between the emotions and consensus judgements from our sample of participants. Our final product was a test of similar length to the original that contains multiracial, age-diverse, and non-gender specific faces across a range of facial expressions.

**Measures.** Apart from the RMET and MRMET, both of which were previously described, we used as our primary measures (1) a 5-item Vocabulary (VOC) test asks participants to choose the closest synonym from among 5 options, (2) the 90-second Digit Symbol Coding (DSC) task that is conceptually similar to the Wechsler Adult Intelligence Scale’s Digit Symbol Coding Test, and (3) a Autism Spectrum Questionnaire (ASQ) that consists of a subset of 8 communication-related questions from the widely-used Autism Spectrum Quotient questionnaire. MRMET, RMET, and VOC are scored as percent correct. DSC is scored as the rate of correct responses (correct responses per minute). ASQ is scored as the cumulative sum of a likert scale self-rating (after reversing the coding of questions that are worded in the opposite direction). We focus on accuracy (either number correct or proportion correct) as the primary outcome measure or score for ASQ, VOC, RMET, and MRMET. The RMET, ASQ, DSC, and VOC tasks are further described in the RDoC Report (Passell et al., 2019). Finally, self-reported demographic variables (age, sex, and ethnicity) of the participants were also recorded, though providing this information was not required.

Because understanding the tests’ internal reliability enables calculation of maximum possible correlations between any two measures, we assessed the internal reliability of each test that we correlated with either RMET or MRMET in our study. For Digit Symbol Coding (DSC), in order to assess both accuracy and reaction time, we assessed internal reliability by comparing the average reaction time until the next even or odd question correct (efficiency
score). Correlation between odd and even halves of DSC was very high (Pearson’s $r = .85$), with an adjusted split half reliability of .90. For the shortened ASQ and the 5-item Vocabulary tasks, we assessed Cronbach’s alpha rather than split-half reliability due to the small number of questions asked within each study. Cronbach’s alpha for ASQ was .54 (95% CI: [.49, .58]), and Cronbach’s alpha for VOC was 0.37 (95% CI: [0.3, 0.43]). The relatively low internal reliability for VOC was due, in large part, to its extremely short length: reliability increases with number of items or test length. Additionally, most participants correctly reported the answer for one of the five question items, which may have further decreased the internal reliability.

**Participant Recruitment and Selection Criteria.** To enable contextual interpretation of MRMET scores from a reference distribution, we collected a large normative dataset that includes information for age, gender, education, and ethnicity of each participant. All participants were recruited voluntarily online through TestMyBrain.org, an internet-based research platform through which participants are voluntarily recruited to complete each a series of tests (“battery”), yielding comparable results as from an in-person laboratory while accessing a diverse pool of participants (Germine et al., 2012). We introduce here our normative datasets ($N = 8,060$ and $N = 17,319$) that are used to characterize the distribution of MRMET and RMET scores, respectively, across these demographic categories. We provide the full normative data set in an open-source format to enable flexible comparisons in future research.

Next, in order to observe the relationship between RMET and MRMET — with respect to interchangeability, demographic effects, and validity — we administered a battery of tests to two separate experimental samples consisting of two testing paradigms. These participants, also recruited through TestMyBrain.org in a manner similar to the normative samples, maintained a sample size of $N = 811$ for the half-test paradigm and $N = 1,156$ (Table 1.3; for a description of
the paradigms, see caption). The demographic information of all participants recruited for this study is noted in Table 1.3.

Table 1.3: Summary of demographic characteristics of the four normative and experimental datasets utilized in the study.

<table>
<thead>
<tr>
<th>Demographic Characteristics</th>
<th>RMET normative sample</th>
<th>MRMET normative sample</th>
<th>Half-test paradigm sample</th>
<th>Full-test paradigm sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (% female)</td>
<td>56.4%</td>
<td>56.2%</td>
<td>64.2%</td>
<td>59.8%</td>
</tr>
<tr>
<td>Age (25th / 50th / 75th percentile)</td>
<td>19 / 24 / 35</td>
<td>18 / 24 / 36</td>
<td>18 / 23 / 34</td>
<td>18 / 23 / 33</td>
</tr>
<tr>
<td>Education (% of participants of age ≥ 25 with at most a high school degree)</td>
<td>15.8% (1287 / 8138)</td>
<td>13.1% (393 / 3,006)</td>
<td>11.0% (232 / 2116)</td>
<td>12.4% (220/1775)</td>
</tr>
<tr>
<td>Ethnicity (% European / part-European / non-European of those reporting)</td>
<td>71.1% / 0.1% / 28.8%</td>
<td>70.4% / 5.0% / 24.5%</td>
<td>68.6% / 5.6% / 25.3%</td>
<td>70.4% / 5.2% / 24.1%</td>
</tr>
<tr>
<td>Sample size (N)</td>
<td>17,319</td>
<td>8,060</td>
<td>811</td>
<td>1,156</td>
</tr>
</tbody>
</table>

Experimental samples, which are further described in Procedure and Test Battery, either consisted of participants who took half of both RMET and MRMET (“half-test paradigm”) or just either one of the full RMET or MRMET (“full-test paradigm”). Part-European ethnicities are defined as those who self-report as being European in addition to another non-European ethnicity. The ability to identify as multiple ethnicities were available only for MRMET normative samples, half-test paradigm participants, and full-test paradigm participants.

All participants between ages 12 to 89 who completed all tests without repeats were included, barring those who presented with significant technical difficulties. Those who did not complete the full battery or reported having taken one or more of its tests before were excluded from analysis.
Procedure and Test Battery. A series of tests were given to a large sample of persons to assess correlations of tests of interest with other measures. In this paper, we analyze two separate batteries that were provided to participants. One battery, used for analyzing the relationship of either the full RMET or MRMET with autism-related traits (convergent validity) and a unrelated task for cognitive speed and symbol matching (divergent validity) consisted of the following tests: either the full RMET or MRMET, randomized, shortened Autism Spectrum Questionnaire (ASQ), and Digit Symbol Coding (DSC), in this order (N = 1,156). The second battery, used for analyzing the relationship between the RMET and MRMET in addition to others, consisted of the RMET (Baron-Cohen 2001), MRMET, DSC, shortened ASQ, and 5-item VOC (Vocab), in this order (N = 811). For this battery, the two RMETs were truncated into halves for test length in order to keep the length of the overall RMET-style test the same as the original while also testing all participants on at least a portion of both tests. Specifically, we randomized participants to take either (A) the first half of the RMET followed by the second half of the MRMET or (B) the first half of the MRMET followed by the second half of the RMET.

Data analysis. Data analysis was performed using the R statistical language. Internal reliability was computed using either Spearman-Brown attenuation-corrected internal reliability ($\rho_{xx}$, where x and x' indicate first and second halves of each test), or Cronbach’s alpha ($\alpha$). Pearson’s correlation ($r$) was used for computing correlation between numeric variables, and Spearman’s rank-order correlation ($\rho$) for ordinal variables (education levels). When averaging $r$ values between first and second halves of one test and full version of another test, we first converted $r$ values of each half using a Fisher’s $z$ transformation; then, we took the average between these two terms. We then converted the $z$ values to $r$ values. Maximum correlations between two tests were computed using geometric means of the internal reliabilities.
Differences between groups are analyzed using the effect size measure Cohen’s $d$, with 95% confidence intervals.
Results

In order for the MRMET to be a viable replacement for the RMET within a diverse population, MRMET must show that it is (1) interchangeable with RMET in what it measures, (2) have a normative distribution that is comparable to the RMET across the whole population as well as individual demographic groups, and (3) behave similarly to the RMET with both tasks correlate well with RMET and tasks that correlate little with RMET. If MRMET outperforms RMET in any of these ways, then such a result would suggest an added utility of MRMET over RMET in certain scenarios. In this section, we will address each of these requirements for MRMET to be a reliable and valid replacement for the RMET in a diverse population.

Interchangeability: MRMET captures the same signal as RMET.

To what degree do the reliable and valid signals of RMET depend on the face stimuli chosen by the original test designers? An effort such as ours to produce an alternate version of the test relies squarely on the hope that alternate stimuli can be found that are consistent with our three key aims: racial inclusivity, non-gendered answer choices, and ground-truth-linked correct responses. Ideally, MRMET—even with its entirely new set of stimuli—should remain consistent with these three aims yet fully interchangeable with the existing RMET.

However, we must address the following question: what constitutes interchangeability? The most direct evidence of interchangeability is a high correlation of the new stimuli with the old stimuli. In this context, it is instructive to think in terms of reliably measured variation. Scores on any measure are partly due to reliably measured variation (i.e., “signal”) and partly due to measurement error (i.e. “random variation” or “noise”). Measurement error may include, for example, lucky versus unlucky guessing on forced-choice questions like those in the RMET and MRMET. The higher a test’s measurement error, the lower its reliably measured variation, and
the less it is capable of correlating with another measure. In the extreme, if the reliably measured variation in two measures is exactly the same, then they should correlate with each other to the same degree as they correlate with themselves. That is, the reliability of the tests being correlated constitutes the theoretical ceiling of their correlation with each other, their maximum possible inter-correlation.

Perhaps the most concrete and tangible way to quantify reliability is to look at the correlation between the two halves of a test, which we quantify here. The gray scatterplots in Figure 1.3 show the correlation of RMET’s first half with its second half (bottom-left: \( r(583) = 0.49, 95\% \text{ CI } [0.41, 0.55], \rho_{xx'} = 0.66 \)) and the correlation of MRMET’s first half with its second half (top-right: \( r(569) = 0.45, 95\% \text{ CI } [0.37, 0.52], \rho_{yy'} = 0.62 \)). Notably, these two correlations were statistically indistinguishable (\( r_{\text{diff}} = 0.04, 95\% \text{ CI } [-0.05, 0.13] \)), suggesting comparable levels of reliability for the two tests. On the other diagonal of Figure 1.3, the white scatterplots show the cross-test correlations between one half of the RMET (e.g. first half) and the other half of the MRMET (e.g. second half). Notably, these two correlations were similar in size to the within-test correlations of RMET with RMET or MRMET with MRMET. The correlation of RMET’s first half with MRMET’s second half was \( r(583) = 0.43 (95\% \text{ CI } [0.34, 0.50]) \), and the correlation of MRMET’s first half with RMET’s second half was \( r(583) = 0.44 (95\% \text{ CI } [0.35, 0.51]) \). These two cross-test correlations are not only numerically similar to, but also statistically indistinguishable from, the two within-test correlations. That the cross-test correlations come this close to exactly matching within-test correlations demonstrates essential equivalence between the RMET and the MRMET. That is, MRMET scores capture all that RMET scores capture. This extraordinary degree of convergence between the two tests suggests that their scores are fully interchangeable.
Such interchangeability is surprising. It could easily have been that there was something special about the particular stimuli chosen for the RMET that was difficult or impossible to replicate. Alternatively, even if it were possible to replicate the signal in the RMET with new stimuli, it could easily have been that the three additional requirements we imposed for the MRMET — racial inclusivity, non-gendered answer choices, and ground-truth-linked correct responses — could have imposed limitations on the potential for interchangeability. In the
worst-case, there could have been hard tradeoffs between one or more of MRMET’s requirements and fully valid measurement of mental state referencing or social cognitive ability. In this case, there might have been little to no correlation between RMET and MRMET. However, this was not the case. These two measures appear to be fully interchangeable, which is the first indication that the MRMET — with the goal of diversity — may serve as a stepping stone for normalizing multiracial test development from a homogeneous test.

Sweet spot: MRMET shows more optimal average performance

We characterized the overall score distributions of the RMET and MRMET. We found that participants, on average, scored 9.4 percentage points higher on the RMET than on the MRMET. For the RMET, a left-skewed score distribution was present (mean = 70.6%, median = 73.0%), implying the presence of a mild ceiling effect (Figure 1.1). On the contrary, the MRMET yielded a normal score distribution (mean = 61.2%, median = 61.1%; Figure 1.1) centered almost precisely near midway between chance (25%) and perfect (100%) score, or 62.5%, which allows for maximal precision in normative comparisons (Wilmer et al., 2013). The differences in score distribution suggests that MRMET, though higher in difficulty than RMET, yields an approximately normal score distribution that is maximally positioned to provide precise comparisons of variability in emotion recognition ability.

Demographic mirroring: MRMET and RMET show the same demographic patterns.

While we have now shown interchangeability between RMET and MRMET in the most direct sense, peering into the relationship between these two tests and demographic variables may show whether these tests perform similarly across diverse populations, If our goal is to
create an inclusive test, it is necessary to ensure that MRMET does not introduce new biases or abject discrimination into emotion recognition assessment.

To that end, we assessed score distributions across demographic characteristics. Four demographic characteristics of interest were similarly associated with performance on the RMET and MRMET. First, female participants outscored male participants to a small but similar extent on both the original RMET (Cohen’s $d = 0.34$, 95% CI: [0.30, 0.39]) and the MRMET (Cohen’s $d = 0.33$, 95% CI: [0.30, 0.36]; **Figure 1.4A**). Second, European participants scored higher than non-European participants on both the original RMET (Cohen’s $d = 0.82$, 95% CI: [0.79, 0.86]) and the MRMET (Cohen’s $d = 0.70$, 95% CI: [0.64, 0.75]; **Figure 1.4B**). Third, based on Spearman’s rank-order correlations for middle, high, college, and graduate school-educated samples, increased education was associated with slightly better performance on both the RMET ($\rho = .18$, 95% CI [.17, .21]) and MRMET ($\rho = .19$, 95% CI [.16, .22]; **Figure 1.4C**). Finally, increased age was, on average, associated with better performance on both versions of the test until middle age (**Figure 1.4D**). We note that the normative dataset for the original RMET, provided for the sole purpose of analyzing demographic trends, incorporated a slightly different answering format (drop-down list) from what was incorporated in other sections of this study (button click). Despite the slight differences in average scores in the two original RMET versions, we still observe that the RMET and MRMET maintain similar patterns of score distributions across demographic conditions.
Figure 1.4. Patterns of score distributions across demographic subgroups are similar across both original RMET and MRMET. (A) Scores from gender subgroups across male, female, and other for RMET or female, nonbinary, and male groups for MRMET. (B) Scores from ethnicity subgroups, across non-European and European ancestry for RMET and non-European, part-European, and European for MRMET. (C) Scores from education subgroups, from individuals completing elementary school to graduate schools. Different subgroup labels are a result of different questionnaires given across the normative dataset collection for RMET and MRMET. (D) Age curve demonstrating trends across the lifespan for scores on the RMET and MRMET. Gray region denotes standard error of the moving average (spline). Red vertical lines above and below the red dot (mean) in figures A, B, and C indicate 95% confidence intervals.
Social link: MRMET and RMET correlate similarly with self-reported autistic traits.

The RMET is expected to correlate well with assessments testing similar constructs, also known as convergent validity. A landmark test for assessing autism traits, the RMET has been shown to be correlated with autism phenotypes (Baribeau et al., 2019; Peñuelas-Calvo et al., 2019). As such, it is of particular importance that any new measure developed to reproduce results from the RMET correlates equally as well with autism traits. To assess whether the MRMET is well-associated with autism traits, we computed the correlation of each half of the RMET with the ASQ, as well as the correlation of each half of the MRMET with the ASQ, resulting in four correlations (Figure 1.5). This two-halves approach allowed us to test the reproducibility of the association of the ASQ with each measure. Based on this analysis, the ASQ was similarly associated with both halves of the RMET (average $r = 0.14$, 95% CI [.08, .20]) and both halves of the MRMET (average $r = 0.13$, 95% CI [.07, .20]). Given the internal reliability of the half-test versions of both RMET and MRMET, the maximum observable correlation between each half of RMET or MRMET with ASQ is approximately .50, explaining the low $r$ values associated with both tests and the ASQ. The adjusted correlations for each test, averaged across both halves, was $r = 0.24$ (95% CI [.14, .35]) for RMET and $r = 0.28$ (95% CI [.12, .33]) for MRMET.
Figure 1.5. MRMET and RMET correlate similarly to self-reported autistic traits. Both the full RMET (top) and full MRMET (bottom) were correlated with the full ASQ, with two separate halves of each RMET or MRMET serving as replication for correlational analysis. Pearson’s $r$, sample size, and confidence intervals are denoted at the top left corner of each panel. All halves of all tests show similar correlations with ASQ.

**MRMET and RMET show similar patterns of correlations across vocabulary and digit symbol coding ability.**

Finally, as RMET is expected to associate less strongly with assessments testing unrelated constructs (divergent validity), we assessed whether MRMET and RMET were similarly correlated with tests that measure an ability that is entirely different from what RMET or MRMET aim to measure. One such example was DSC, which is a symbol-to-digit matching
task. Another example is VOC, which is a test of one’s vocabulary knowledge. We computed correlations between scores on MRMET or RMET with scores on both of these assessments. Taking the average between the first and second halves of either RMET or MRMET and adjusting for maximum correlation, both the RMET and MRMET were more strongly associated with VOC (MRMET adjusted $r = .46$, 95% CI: [.24, .67], RMET adjusted $r = .64$, 95% CI: [.44, .84]) than with DSC (MRMET adjusted $r = .30$, 95% CI [.16, .43]; RMET adjusted $r = .36$, 95% CI [.23, .49]; Figure 1.6). Furthermore, the association between the RMET and VOC was slightly stronger than the association between MRMET and VOC.

Taken together, MRMET shows both a similar relationship with social outcomes and similar reliability as does RMET, all the while maintaining a potentially lower but not statistically different association with VOC performance. As RMET is not intended to be a vocabulary test, the lower association between MRMET and VOC is indicative that MRMET is measuring a specific behavior rather than one’s general vocabulary knowledge. Thus, MRMET maintains similar convergent validity of RMET while improving on its divergent validity.

**MRMET shares similar construct and predictive validity as the RMET.**

*Figure 1.6: Construct Validity, Predictive Validity, and Vocabulary Loading.* Pearson’s correlation, $r$, with 95% confidence intervals across all halves of both RMET and MRMET with ASQ, DSC, and VOC. All $r$ values and CIs shown are attenuation-corrected except RMET H1/MRMET H2 and RMET H2/MRMET H1 under Construct Validity, which are unadjusted.
Discussion

We have just described in detail the process of developing, assessing, and validating the Multiracial Reading the Mind in the Eyes test in direct comparison with the Caucasian Reading the Mind in the Eyes Test. Our analysis of the MRMET, across a large and diverse sample, shows that it has the same or similar psychometric and characteristics across demographics as the original RMET. In summary, the two tests had similar split-half reliability, with correlations in scores across both tests the same or similar to the split-half reliability. Scores on the two tests had similar associations with demographic characteristics such as age, race and ethnicity, gender, and education. Associations with scores on the Autism Spectrum Quotient were similar across both tests. Divergent validity (based on correlations with processing speed based on digit symbol matching performance) was also similar across both the original RMET and the MRMET. MRMET performance was less associated with vocabulary performance, however, compared to RMET. Thus, the only notable psychometric difference between the two tests is a potentially lower reliance on vocabulary for the MRMET.

From a validation standpoint, the RMET and the MRMET are likely measuring the same underlying construct, with a similar level of reliability. The correlation in performance between the MRMET and the RMET was at the ceiling for possible correlations based on the internal reliability of each test. Thus, the MRMET could be used in place of the RMET in many contexts—particularly those where the target population of participants is not homogeneously of European ancestry. While the RMET was designed for use in a relatively homogeneous — though, still, not entirely homogeneous — population based in the UK (Baron-Cohen et al., 2001), it has now become one of the most widely used measures of social cognition across the world. The success of the RMET thus became the reason for one of its most critical failings—a stimulus set that doesn’t represent the populations that the RMET is used to assess.
(Dodell-Feder et al., 2020). The development of the MRMET is our attempt to keep all that is useful about the RMET, but with stimuli that reflect the diversity of participants that the RMET is now used to assess.

Although the MRMET employs a much more diverse set of stimuli than the original RMET, both tests had a similar pattern of association with race and ethnicity. Like in the original RMET (Dodell-Feder et al., 2020), MRMET scores were higher for participants of European descent than participants of non-European descent. This would indicate that the MRMET still contains biases related to race that are not addressed by the use of multiracial faces.

While stimuli were taken from videos of actors who were instructed to express the indicated complex emotion, items were selected based on convergence between these emotions and consensus judgements from our original sample of participants. As this sample was 75% participants of European descent (¾ of European participants were from the United States), this means our stimulus selection process was biased towards European-centric and US-centric emotion interpretations (Elfenbein & Ambady, 2002).

While research has indicated that there are basic expressions of emotion that are shared across cultures (Ekman & Davidson, 1994), there is also significant evidence of sociocultural variability in the expression and interpretation of emotions (Barrett & Russell, 2014). This means that—to some degree—the accuracy of mental state understanding will depend on a match between the participant’s background of social learning, the background of social learning of the actor, and the background of social learning of the participants from whom consensus answers are derived. Where this source of bias in the RMET and MRMET comes from is an open question that is worthy of further research. We note, however, that it is not possible to target both high convergence with the RMET and address all the generalizability limitations of that test. From a psychometric standpoint, a multiracial test cannot be used as an alternative to the
RMET and show a different pattern of association with key demographic and individual differences variables. Our goal here was to produce a culturally appropriate and more racially representative alternative to the RMET, recognizing that we have not achieved a fully generalizable measure of social cognition that avoids all the biases in the original measure. While we believe our development and validation of the MRMET is an important step, more work is needed to understand how (and whether) we can measure social cognition in a broadly generalizable and unbiased fashion.

In addition to providing a substitute for the RMET in studies with non-European participants, the MRMET could also be used as an alternate form—for example, in studies where multiple measures of mental state inferencing is desirable. Currently, there are no parallel forms of the RMET to facilitate its inclusion in trials and longitudinal studies. While we did not look at test-retest reliabilities of the RMET and MRMET in this study, alternate forms reliability (or correlation between two forms of the test) was as high as could be expected given each test's internal reliability.

One of the reasons the original RMET gained such broad use is the laudable choice of the original creators to make the test widely available through the website of the University of Cambridge Autism Research Centre. Although broad distribution of stimuli might result in the overexposure of those stimuli, which limits their use in diagnostic and research settings, we believe that the advantages of accessibility and open science outweigh the disadvantages. Thus, we have created a package of materials distributed under an open source license (CC-BY-SA) that will allow others to use the stimuli and test in their research studies. Interested researchers may also contact us to obtain access or integrate web-based implementations of the MRMET into their studies.
There are several limitations of the MRMET and our current validation work that should be considered. First, data collection was done entirely over the web. While we have previously found that web-based and lab-based administrations of the RMET have similar psychometric characteristics (Germine et al., 2012), it is possible that factors related to the context of administration might impact scores. For example, participants might be more likely to look up the definitions of words in an unsupervised web-based setting than an in-person setting, making the reliance on vocabulary ability higher for an in-person assessment. We also note that our analyses of the associations between test performance and sociodemographic characteristics were based on data collected over different time periods (data on convergent and divergent validity was collected at the same time). Thus, it is possible that patterns were distinct between the RMET and MRMET, but that these differences were offset by population-related changes in the association between performance and sociodemographic variables over time, making the two tests appear more similar. Finally, as noted above, although the MRMET contains a diverse range of face stimuli, there is still a performance advantage for participants who self-reported European ancestry. Such differences may be related to the way social and cultural factors that vary with race and ethnicity influence the interpretation of facial expressions (Elfenbein & Ambady, 2003).
Conclusion

Our science can only be as good as what we measure. The representation and inclusion of racial diversity in behavioral science is an ethical imperative, and one that should be reflected in our research measures. The study of social cognition—particularly as it relates to cognitive and mental health—cannot exclude those populations that suffer from the greatest impacts of health disparities. If we believe that understanding social cognition is important for understanding mental health, then our near exclusive reliance on Caucasian stimuli in social cognition research is both scientifically and ethically unjustifiable. The continued reliance on such stimuli violates trust, reduces the generalizability of our science, and—based on feedback from our own participants—constitutes a microaggression against communities of color who are already marginalized in science and health research. We consider the development of more inclusive and socially responsible alternatives to existing measures to be both a scientific and ethical imperative if we are to achieve a robust and generalizable science of social cognition and behavior. We hope that the development and validation of the MRMET—as a multiracial alternative to the classic Reading the Mind in the Eyes Test—leads us a step closer to that goal.
Chapter 2: Primer for Behavioral Genetics and Genotype-Phenotype Analysis

Introduction

In Chapter 1, I investigated the relationship between two alternate behavioral measurements of mental state inferencing: the Reading the Mind in the Eyes Test (RMET) and Multiracial Reading the Mind in the Eyes Test (MRMET). But, where does behavior come from? Notably, behavior is—in part—shaped by genes. This fact has been studied in a variety of ways: neuroscientists, for example, have created genetic variants in order to study how such a variant causes behavioral alterations in a laboratory animal. Psychologists have studied the behaviors of non-identical twins (who share approximately half of their genetic variants with each other) and study how they might differ from identical twins (who share all genetic variants with each other). More recently, large genomic databases that pair genetic information with behavioral information have become widely available, which has enabled researchers to systematically study the causal relationships between genes, their variants, and phenotypes. All of these approaches have successfully shown that genes do share some part in explaining behavior.

In Chapter 2, I focus on genomic research. As the technology involved in making and analyzing very large databases improves, we are able to gather large-scale information on both genotypes and phenotypes. Through such databases, researchers have begun to understand the detailed relationships between one’s genetic code and associated traits. Such an approach has led to fruitful innovations, including therapeutics for cardiovascular disease and the ever-nearing future of precision medicine. Furthermore, large genomic data has the amount and depth of information required to describe, in biological detail, specific biological correlates or
causes of behaviors and traits. For example, in autism spectrum disorder (ASD) research, there are various databases like SPARK, the Danish iPSYCH repository, and Simon Simplex Collection (SSC) that provide both genetic data for every participating individual and detailed behavioral testing results. By deeply annotating each individual’s genetic information with behavioral information, we can come to understand the ways that natural genetic variation may contribute to making every individual different from each other. Due to the depth and extent of data collection, the data aggregation approaches of the aforementioned databases are particularly helpful in understanding the biological architecture of very complex disorders like ASD.

In this intersection of biology and behavior, good research requires both good tests for behavioral data and sophisticated algorithms for genetic data. In Chapter 1, we validated an inclusive, reliable, and valid behavioral test. In this chapter, I now segue into the conceptual framework that has recently gained popularity for describing the impact of many thousands of genetic variants on a given phenotype (that is, a given behavior). This concept, called the polygenic risk score (PRS), is a nudge towards the idea of dimensionality in psychiatric conditions. The process for calculating PRS for an individual relies on computation, statistical tools, and recent discoveries in genomics. To detail the scientific rationale of PRS, I will first outline in detail the concept of a biological spectrum as it pertains to ASD. Here, I will describe what is currently known about the genetic architecture and polygenicity—i.e., the effect of multiple genes—of ASD and schizophrenia (SCZ). Second, I will describe the scientific intuition for creating simplified values from an individual’s entire genome. Lastly, I will mention the methods and softwares that were employed in this instance to calculate the PRS scores. Soon, in Chapter 3, I will apply the concepts discussed here in order to directly analyze the relationship between clinically-derived PRS scores and a set of behavioral measures (a survey
and an interview), using data from the SSC database, which will seek to shed light on the biological architecture of a particular set of clinically-relevant complex behaviors.

Biological Spectrum in ASD and SCZ

Common traits in people that present themselves in diverse ways are often the result of multiple genes. Take, for example, one physical example: height. There are many genes that each affect a person’s height to a different extent (i.e. have varying “effect sizes”). Many complex traits share this biological architecture. People have different genetic variants, and the overall effects of individually subtle genetic variation contribute to large-scale human variation.

Does this concept apply to behaviors? Indeed, many human behaviors—though some much more than others—are shaped by genes and their variants. The extent of biological influence on human behavior is extensive: In one example of how common genetic variation influences common behavior, the degree of genetic effect on personality traits was estimated to be about 40% (Vukasović & Bratko, 2015). In a similar manner, possessing many particular common genetic variants can also predict one’s risk for a psychiatric diagnosis for ASD or SCZ because the number of risk variants for ASD or SCZ one harbors increases one’s liability for the condition. Together, this information shows that the additive effects of natural human variation, at a certain threshold, confers greater liability for a particular trait, personality, or psychiatric condition (Felsenstein, 2012; Guloksuz et al., 2019; McIntosh et al., 2006).

The genetic liability model enables us to biologically conceptualize the autism spectrum disorder and schizophrenia across a broad spectrum. Over time, the perception of these two conditions have become increasingly dimensional: both present themselves in varied symptoms among the clinically diagnosed population, and the conditions themselves are highly heritable. Furthermore, many associated genetic risk variants for ASD and SCZ have been identified, with
significant overlap between the two conditions (Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013; Kushima et al., 2018). Therefore, in comparison to most other psychiatric conditions, ASD and SCZ are relatively unique in the way they are diagnosed: while a clinician can arrive at many major psychiatric diagnoses through “yes”-or-“no” categorical checkboxes reliant on the DSM-V, the diagnosis of ASD and SCZ have underlying variability of symptoms that is baked into the very definition of what constitutes the condition.

**Autism spectrum disorder.** ASD is a developmental condition that emerges during childhood and is characterized by broad symptoms, including social and communication impairments and restricted and repetitive interests, among other symptoms. Autism was first described by Leo Kanner through personal observations about 80 years ago, but our knowledge about the condition has evolved considerably since then. ASD commonly (~35%) presents with intellectual disability (Srivastava & Schwartz, 2014), but due to the large variation in symptoms, ASD is now conceived as a spectrum with many possible ways that a particular “case” might look like. A variety of neurocognitive assessments and clinical tools are aimed at understanding this wide individual variation, which have aided in the behavioral characterization of ASD. The tools studied in this thesis — Reading the Mind in the Test, Autism Diagnostic Interview-Revised, and the Social Responsiveness Scale — are but a few of many diagnostic tools for studying or identifying autism. Genetically, twin studies suggest a large genetic component to ASD risk, with a heritability of approximately 60-80%. Yet, as for most complex traits, genomic investigations have shown that there is no single causal gene for ASD. To identify the specific genetic variants involved in autism risk, more recent studies have focused on both common and rare genetic variation, including 71 genes associated with ASD through large-effect rare variation (Satterstrom et al., 2020), and 5 common variants of smaller effect (Grove et al., 2019).
Schizophrenia. SCZ was first described over a century ago by Bleuler, who relied on largely subjective observations, as was the norm at the time (Maatz et al., 2015). However, in more recent years, SCZ is diagnosed based on a group of common “positive” symptoms (such as hallucinations, paranoia, or delusions), and common “negative” symptoms (such as catatonia, lack of interest, and withdrawal), and cognitive dysfunction. It is now known that there are no obvious anatomical anomalies in the brains of those with SCZ (Birnbaum & Weinberger, 2017). Genetically, many schizophrenia-associated genetic variants are responsible for early neurodevelopmental processes and it is believed that early brain development mediates some of the genetic risk of schizophrenia (Birnbaum & Weinberger, 2017). Furthermore, SCZ, like ASD, is heritable, with heritability estimates as high as ~80% (Hilker et al., 2018).

How were the abovementioned genes discovered? There are many possible ways to discover a gene that is known to affect a trait, such as observing how a genetic knockout strain of a laboratory animal behaves differently from a normal one, or taking a case study of a person who is notable for a particular genetic anomaly. Yet, another method that relies on large-scale, highly exploratory, and multiple human-genome based information is the genome-wide association study.

Genome-Wide Association Studies

A genome-wide association study (GWAS) is one method that enables the discovery of a causal relationship between a genetic variant and a characteristic. A GWAS is able to simultaneously examine a multitude of genetic variants within a population and observe how they each relate to a trait of interest. GWAS studies build upon (1) the Human Genome Project, which has enabled the sequencing of variants, and (2) the International HapMap, which identified approximately 4 million locations of single nucleotide polymorphisms (SNPs) as well
as association between SNPs. SNPs are a location in the genome that has a common single-nucleotide difference across the population that is indicative of individual variation. Most SNPs exist outside of coding genes as only 1-2% of the genome is composed of genes; such SNPs are considered non-coding variants (Zhang & Lupski, 2015). Any association between SNPs represents how often any two SNPs tend to co-occur in a person. This is also known as linkage disequilibrium or LD (Chang et al., 2018). Through these animating developments in genomic research, we can now investigate the relationship between SNPs commonly found in the general population (“common variants”) and how any given SNP affects a particular trait with statistical accountability. This process therefore encapsulates the search for causal genetic variants for a trait, even if the effect of any particular genetic variant on a trait is small.

In order to conduct GWAS, one must first obtain samples from the population who show the particular trait of interest (“cases”) as well as samples from the population who do not show the particular trait of interest (“controls”). One common method of finding such samples is through random sampling of the population, but an alternative to this approach includes the recruitment of entire family units. Once the samples, which must be very large, are obtained, the subjects’ DNA is tested for common SNPs using chip-based microarrays. These microarrays, often through Affymetrix or Illumina platforms, survey the genome for millions of common SNPs and yield the genotypes for each individual.

Upon gathering genotype information, one can apply significance testing (often at a threshold of P-value < 1x10^{-7}) between cases and control groups. Of all the SNPs that show a significant difference, quality control (QC) omits certain samples that do not meet quality checks, such as results of arrays that are inconsistent with expected or known values. Finally, for all resulting SNPs, multiple hypothesis testing correction is applied in order to limit error rate and false discovery rate. Finally, the empirical distribution of the test statistic is computed via
permutation testing. The resulting SNPs are now considered associated variants, but are subject to replication of these results in another sample in order to distinguish between false discoveries and true associations. At this point, these identified SNPs are mapped onto the “susceptibility gene locus,” i.e., the region of the gene identified by the microarrays, by fully sequencing the DNA base pairs in that particular region and identifying the genetic alteration involved in causing the disease. This final step weeds out the discovery of non-causal SNPs (“tag SNPs”) that happen to be inherited frequently with another, truly causal, SNP; most associated SNPs are tag SNPs.

Based on this process of identifying common risk variants, then, how might one biologically characterize the overall risk for a particular condition? Polygenic risk scores (PRSs) capture the notion of an inherent biological spectrum by summarizing multiple variants and their effects into one value that captures how likely it is that the individual demonstrates the particular trait.

**ASD and SCZ Polygenic Risk Score**

Polygenic risk scores (PRS) are computed by aggregating the estimated genetic effects of a variant from GWAS, thereby allowing for the approximation of a *cumulative* effect of common genetic risk on a trait. A PRS aims to compute a single score that measures both the number of common variants—i.e., genetic variants that are commonly seen in the general population—that a person harbors, as well as each variant’s associated, cumulative effects on the phenotype of interest.

Let us now consider two specific conditions that can be studied via PRS: ASD and SCZ. In Chapter 3, we will study the relationship between ASD and SCZ genes with social phenotypes using PRS scores. As one example that showcases the process of PRS calculation,
we provide here the method used in Simon Complex Collection to compute the PRS score. While this is only one example of the many methods of computing PRS, this example is the same method that has been employed in a variety of recently published research. I will focus on the data sources, GWAS results, and computational tools that were employed in the process of computing ASD and SCZ PRS in the Simon Simplex Collection. I will now dive into deeper detail about the protocol for PRS score computation, adapted from Choi et al., 2020 and Fischbach & Lord, 2010.

**Data acquisition.** The first step of the process of computing a PRS for a phenotype often utilizes GWAS results. This information is often acquired through a published GWAS or multiple GWAS (e.g. from databases such as GWAS Catalog). In addition to information on common variants and their effect sizes obtained by GWAS, there must also be a genotyped dataset of sample participants for whom the PRS will be calculated. This is because PRS is computed for a particular person, meaning, a genetic sample independent from those used to find the results of GWAS will be a “test” case of the PRS computation. These requirements thus require the acquisition of data of two kinds: data that describes common genetic variation, and a group of people whose genetic variation we wish to study.

In the PRS score example later utilized in Chapter 3, there are two main GWAS involved for ASD and SCZ. For ASD, GWAS results were sourced from GWAS data from the Denmark iPSYCH Consortium, which sampled and meta-analyzed 18,381 Danish individuals with ASD and 27,969 Danish controls (Grove et al., 2019). These samples yielded five loci in the genome that are associated with ASD. For SCZ, the GWAS dataset results were obtained through the Psychiatric Genomics Consortium (PGC), which collected up to 36,989 cases of SCZ and 113,075 controls. Through these samples, the PGC identified 108 significant loci—83 of which were previously unreported—after multiple hypothesis testing correction (Ripke et al. 2014).
The “target” data — i.e., genomic data of the people for whom PRS is to be computed — was collected through the Simon Simplex Collection (SSC), which was a joint effort between SFARI and 12 other institutions. The SSC collected family-based samples in which one child — of one or multiple total children within the family — had ASD, yielding a total sample size of 1,887. The SSC genotype data was sequenced from the Illumina microarray using a blood sample from each individual.

**Quality Control.** Upon identifying GWAS datasets, in addition to acquiring the target samples’ genetic data, researchers apply quality control (QC) to the acquired datasets in order for the resulting PRS to be valid, accurate, and consistent. One must ensure that both the GWAS and sample data meet criteria for reliable results.

To compute a PRS, one must ensure that the SNPs of interest contribute to variation in the phenotype such that the phenotype has some meaningful SNP heritability. Because GWAS are the main ways through which we determine risk loci, one should only perform PRS for GWAS data with a $h^2_{\text{SNP}} > 0.05$ ($h^2_{\text{SNP}}$ refers to the SNP heritability of the trait). Finally, across both GWAS and target population datasets, SNP quality must be controlled for. This includes removing ambiguous, mismatching, and duplicate SNPs. Overlapping samples between the GWAS data and target data must also be removed in order to avoid spurious association.

**Statistical Adjustments.** Post-QC, there remain statistical inconsistencies within the genome that must be accounted for. One type of statistical assumption that is commonly made is independent assortment; however, some SNPs are commonly inherited together and causes a “linkage” between these two unrelated SNPs, known as linkage disequilibrium (LD). This pairing must be statistically accounted for. If left unaccounted, the SNPs that are unrelated to the trait under study but are frequently inherited together will inflate PRS standard error, a value that is a function of the sum of uncertainties of each SNP. This next step of PRS calculation
accounts for these types of statistical inconsistencies by applying a variety of corrective measures.

To correct for LD, one typically shrinks the effect size estimates of all SNPs. In the PRS computed for the SSC cohort, PRS were adjusted using one particular shrinkage method found in the LDpred software. LDpred is a Bayesian framework for accounting for LD—that is, one must first specify a prior distribution for the genetic architecture and LD information from a reference panel in order to perform shrinkage (Vilhjálmsson et al., 2015). From this specified a priori distribution, the Bayesian framework then updates the distribution to bring the distribution closer to the optimal parameter values for PRS prediction, which includes a parameter for the fraction of causal variants (Choi et al., 2020; Vilhjálmsson et al., 2015).

**Aggregation.** After taking into account the quality of data and statistical errors that might be present, one can sum up the effect sizes of every genetic variant in the target individual. This is the final PRS score. SSC PRS values were generated through PLINK version 1.9. PLINK is a general-purpose genetic analysis software that includes the ability to calculate PRS scores from GWAS by using the flag “—score.”

By utilizing an aggregate score across many common risk variants that can be computed for any particular person, PRS expands the possibilities of gene-trait relationships beyond a genetic study between individuals with a disorder versus individuals without. Rather, it allows us to view people as existing somewhere on a spectrum of genetic liability for a particular condition. To that end, PRS have recently been used in a variety of applications. For instance, Nayar et al. employed ASD PRS as the genetic measure to account in part for a mother’s Broad Autism Phenotype Questionnaire score, which demonstrated that PRS are now strong enough to measure real signals from human behavior. SCZ PRS has been also associated with a variety of behaviors and outcomes in the general population, including creativity (Power et al.,
2015), antipsychotic treatment efficacy (Zhang et al., 2018), and many others (Mistry et al., 2018). These successful applications of PRS with traits in the general population suggest that (1) PRS now captures enough genetic signal that is associated with a particular trait, and (2) PRS holds much utility for investigating genotype-phenotype relationships in the broader population without having to place each person on a dichotomous “yes”-or-“no” category for conditions.

Limitations of Current Approaches

Most genetic research has been focused on either White (~70%) or East Asian (~20%) populations (Duncan et al., 2019). Due to this fact, predictive performance of typical PRS computations has been found to be lower in diverse samples (Duncan et al., 2019). Further, beyond these populations, there has seldom been widespread, ethical data collection that probes at both behavioral and genetic data. Thus, generalizability is a major concern. Especially given the wide-reaching potential of genetic research—such as shaping the future of medical treatment regimes—the pervasive homogeneity in the population studied is unacceptable.

For instance, because GWAS results rely on the populations that they study, the high proportion of Caucasian samples indicate that genetic signals from non-Caucasian populations—which are likely to be different—are suppressed, especially due to the small effect sizes of common causal variants. Thus, GWAS results are likely biased towards discovering associated variants in Caucasian populations compared to non-Caucasian populations. In addition, because PRS scores are directly derived from GWAS results, PRS scores, too, are likely most accurate for those with European ancestry (Duncan et al., 2019). Even more concerningly, models of PRS derived from one population often are not applicable to another:
the common presence of population stratification and the fact that SNPs differ by ancestry together lead to the lack of generalizability of current approaches in modelling disease.

To address some of these pervasive issues, various current efforts aim to expand available samples. One current study, being undertaken by Robinson et al., aims to expand our understanding of behavior and genes beyond the Caucasian and East Asian populations through a project called NeuroDev (de Menil et al., 2019), which is in collaboration with Kenyan and South African universities and local professionals to host a large-scale ethical study of genes and behavior. In another example, Population Architecture using Genomics and Epidemiology (PAGE) conducted GWAS of more than 20 phenotypes on non-European, diverse samples spanning individuals who self-identify as Hispanic/Latino (n = 22,216), African American (n = 17,299), Asian (n = 4,680), Native Hawaiian (n = 3,940), Native American (n = 652) or Other (n = 1,052), with a total sample size of n = 49,839 (Wojcik et al., 2019). These diversity-focused initiatives are critical for developing generalizable PRS and other genetic measures, which, in turn, are the foundation for future medical technologies.

Summary

Polygenic risk scores synthesize a large amount of information. They include many thousands of common SNPs and their respective estimated small effects on a person’s phenotype, and they synthesize the person’s full genetic information into a single number that describes the person’s genetic liability for a particular polygenic trait. PRS thus goes beyond a categorical, yes-or-no approach of determining a person’s trait; it instead places a person on a spectrum of common variants.

However, this promising method—now utilized in many studies that explore genes and behavior—has many limitations. The most concerning limitation, perhaps, is its potentially
limited generalizability beyond the specific sample populations studied. We must recognize that science is only as good as its measures or methods, and its measures or methods are only as good as the samples that were used to develop them. PRS is but one example of this: to the extent that most associated variants discovered through GWAS and most knowledge in genomics come from investigating a mostly European population, we must be careful in applying the insights gained to the broader population.

Expanding on these ideas, I will now examine the genetic correlates, as defined by PRS scores, with a set of survey-based social behavioral measures.
Chapter 3: Genetic Correlates of Social Phenotypes in ASD and SCZ

Background

How is genetic variation captured by the ASD and SCZ PRS related to social behaviors? Surprisingly, this question has not been deeply studied. Past research on ASD PRS and behavior have reported an association between social therapy outcomes (Li et al., 2020), face and emotion recognition ability (Qin et al., 2020; Wendt et al., 2020), and motor and language skills (Takahashi et al., 2020), among others. Other studies have estimated that approximately 30% of common polygenic influences are shared between an ASD diagnosis and autism-like traits in young children in the general population (Robinson et al., 2016). Similarly, as mentioned in Chapter 2, SCZ PRS has been associated with a variety of behaviors and outcomes. However, most of these studies don’t look at the direct relationship between a group of specific social behaviors, such as imaginative play, making friends, or interest in others. Here, I pursue an in-depth analysis of specific behaviors, as measured by diagnostic questionnaires, and their relationships to PRS.

In order to accomplish this, I focus on the relationship between social phenotypes and genetic risk among individuals who have been ascertained for ASD. This approach is different from studying the broad population, where samples are widely available and great phenotypic individual variation exists. By observing only those who have been ascertained for ASD, we can expect that these individuals will have similar scores on gross social ability, meaning that the degree of individual variation is limited due to the niche stratum of the ASD population.
However, despite the pervasive stereotype that ASD-diagnosed individuals are “bad” at all social abilities, it must be recognized that ASD samples are not homogenous. Through this fact, I turn to a more promising approach: I will observe particular characteristics among ASD-diagnosed individuals that are known to vary from person to person. I explore two particular social characteristics within the ASD population that are relatively varied: empathy and prosociality (Altschuler et al., 2018; Bird et al., 2010; Dissanayake et al., 1996; Jones et al., 2010; Ronald et al., 2005). Because there exists individual variation among people with ASD on the extent of experienced empathy and desire to interact with others, looking at the degree of genetic associations for each of these qualities may lead to fruitful insights.

To pursue this study, I must first develop a measure for assessing prosociality and empathy. Thankfully, this work has been recently done by a group of scientists who employed latent analysis on various ASD diagnostic questionnaires. Zheng and colleagues recently discovered four latent factors that underlie the individual questions from Autism Diagnostic Interview-Revised (ADI-R), ADOS, and Social Responsiveness Scale (SRS) (Zheng et al., 2020). Among the discovered four factors was a Peer Interaction and Modification of Behavior (PIMB) with 14 questions and Social Initiation and Affiliation (SIA) with 5 questions, which are further described in Table 3.1. Because these two factors describe behaviors that are similar to empathy and prosociality, respectively, these two factors may serve as good proxies for prosociality and empathy even though they are not perfect equivalents of these behavioral qualities.
Table 3.1: Questionnaire items and their sources for the 5 most associated questions within each factor.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Questionnaire Item</th>
<th>Source</th>
<th>Question #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peer Interaction and Modification of Behavior</td>
<td>Has difficulty relating to peers</td>
<td>SRS</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Is socially awkward</td>
<td>SRS</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>Has difficulty making friends</td>
<td>SRS</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Is awkward in turn-taking</td>
<td>SRS</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Plays appropriately with children</td>
<td>SRS</td>
<td>22</td>
</tr>
<tr>
<td>Social Initiation and Affiliation</td>
<td>Response to approaches of other children</td>
<td>ADI-R</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>Group play with peers</td>
<td>ADI-R</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Interest in children</td>
<td>ADI-R</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>Imaginative play with peers</td>
<td>ADI-R</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>Imitative social play</td>
<td>ADI-R</td>
<td>61</td>
</tr>
</tbody>
</table>

Overall, in this chapter, I aim to measure how PRS relate to the factors PIMB and SIA. I hypothesize that if PRS—i.e., the additive effects of common SNPs and their estimated effects on ASD risk—do not correlate with either of the factors, then there are many potential causes. From a reliability and validity standpoint, (1) PRS scores might need validity or reliability improvements, (2) phenotype measurements may require validity or reliability improvements, or (3) the sample size is too small and thus underpowered to detect a small but non-zero effect. Another possibility, which assumes that our measurements were valid, reliable, and our sample had sufficient power, would be that polygenic risk for ASD, in fact, is not involved in the
variability of the two investigated social phenotypes—empathy (PIMB) and prosociality (SIA)—in people with an ASD diagnosis.

On the other hand, if we observe that the ASD PRS is indeed correlated with both PIMB and SIA, then this result may have the following underlying reasons: (1) PIMB and SIA both measure the same aspect of social behavior, which in part reflect polygenic risk for ASD, or (2) PIMB and SIA both measure different social behaviors, but due to pleiotropy—i.e., the ability of one genetic variable to have numerous phenotypic effects—ASD PRS exerts influence on these partially distinct behaviors.

Lastly, if PRS are correlated with only one of PIMB or SIA but not both, then this would be evidence that the factors are each measuring—in part—distinct qualities, and the exact nature of this distinctiveness, such as what causes one factor to be more correlated with the ASD PRS than the other, may warrant further investigation.

Methods

**Subjects.** Data was obtained from the Simon Simplex Collection (SSC), which provides genotypic and phenotypic data for each individual in the sample (Fischbach & Lord, 2010). The SSC, which is operated jointly through SFARI and 12 research institution-affiliated clinics, identified and collected data for SSC participants until 2011. Inclusion criteria for the SSC include (1) a clinical diagnosis of ASD and (2) meeting ASD or autism cut-offs on the ADI-R and ADOS. The SSC data contains information on a group of simplex families—i.e., families in which only one individual, the child (“proband”), is clinically diagnosed with ASD. 1,529 out of 1,887 families studied (“quads”) had at least one unaffected sibling in addition to the proband; the rest of the families (“trios”) had only one child—the proband—and no siblings. Details about the
breakdown of trios and quads in the SSC sample are shown in Table 3.2. Exclusion criteria are further described in Fischbach & Lord, 2010.

In total, the number of families assessed for SSC who passed QC for our study was \( n = 1,462 \), among which samples of European ancestry had a size of \( n = 1,191 \) and non-European ancestry had a size of \( n = 271 \). Because PRS from people of different ancestries tend to behave differently, we split the analysis into (1) all individuals from the SSC cohort, (2) European samples only, and (3) non-European samples only. It is typical to only analyze individuals with European ancestry in projects like this one, but we were interested in observing how PRS scores behave in European samples in comparison to more diverse populations.

Table 3.2: Simon Simplex Collection family samples.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>One Male Sibling</th>
<th>One Female Sibling</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Proband</strong></td>
<td><strong>Quads</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td><strong>Trios (No Sibling)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>306</td>
<td>617</td>
<td>708</td>
<td>1631</td>
</tr>
<tr>
<td>Female</td>
<td>52</td>
<td>92</td>
<td>112</td>
<td>256</td>
</tr>
<tr>
<td>Total</td>
<td>358</td>
<td>709</td>
<td>820</td>
<td>1887</td>
</tr>
</tbody>
</table>

Trios indicate families with one child with autism, whereas quads indicate families with two children, one with autism and one without. Table adapted from Fischbach & Lord 2010.

**Genetic characterization.** From each participant, a blood sample was collected, from which DNA was extracted, cell lines established from transformed lymphoblasts, and analyzed by Rutgers University Cell and DNA Repository. Whole genome genotyping was performed based on NimbleGen chips and on Illumina chips.
**Behavioral Measures.** For behavioral measures, the Autism Diagnostic Interview-Revised (ADI-R), Autism Diagnostic Observation Schedule-2 (ADOS-2), and Social Responsiveness Scale (SRS) were used. The ADI-R is a parent interview questionnaire designed for use by an investigator. It collects information about social communication, restricted and/or repetitive behaviors, and other types of behaviors typically present in a person diagnosed with ASD (C. Lord et al., 1994). While the ADI-R typically collects information on the degree to which the child has displayed the behavior in (1) the past 3 months and (2) at any time during the past, we only analyzed ADI-R scores regarding the entire past. The ADOS-2 is a standardized and semi-structured observational assessment conducted by a clinician for diagnosing a child for ASD. There are five “modules,” each of which is designed for differing levels of expressive language and age; the module that is most appropriate for any given child is selected for use. In this paper, we only include participants administered Module 3, designed for children with fluent speech (C. Lord et al., 2000). Lastly, the SRS is a questionnaire administered to parents and teachers of the subject that measures autism-related symptoms on a scale from “not true” to “almost always true.” In this study, we focus only on parent-reported values.

**Social Behavior Factor Analysis.** From the questions in these three diagnostic questionnaires, Zheng et al. 2020 determined four underlying factors through latent factor analysis. In essence, each of the four underlying factors of the questions are determined by observing which questions correlate best with each other. From this, Zheng et al. determined four clusters of questions that are most related to each other, and each of these clusters were labelled by a descriptive name. Among the four factors described by Zheng et al., we employ two factors: Peer Interaction and Modification of Behavior (PIMB) and Social Initiation and
Affiliation (SIA). From each of these factors, we selected the top 5 items with the greatest correlation within each factor. The rank of items that correlate best within the factor can be found in Zheng et al., 2020. This process yielded 10 questions in total to analyze for this study. Interestingly, all top 5 questions for the PIMB were from the SRS, and all top 5 questions for the SIA were from the ADI-R. The selected questions are included in Table 3.1. All responses to these ten questions used in this study were rated on an integer scale from 0 to 3, where 0 indicates lesser display of social behavior and 3 indicates greater display of social behavior. This scaling is used regardless of the directionality of the question asked (i.e., the same child will have the same score, 3 for more social behavior and 0 for less social behavior, regardless of whether “does your child show social behavior” or “does your child not show social behaviors” was asked).

**Data Analysis.** All data was analyzed in the R statistical language and visualized in R or ShowMyData.org. All correlational values are reported by Pearson’s correlation values, and CI and significance testing between correlational values were computed through the cor.test function in R. For the correlation test, the test statistic is based on Pearson’s correlation (r), which follows a t-distribution with n - 2 degrees of freedom (if the samples follow independent normal distributions). An asymptotic confidence interval, based on Fisher’s r-to-Z transform, was also computed. Based on Fisher’s Z transform of correlation values, I also computed average correlation values as well as the difference between correlations using Vassar’s r-diff tool (http://vassarstats.net/rdiff.html). For genetic analysis, we use the pre-computed proband PRS to determine each individual’s estimated genetic liability for ASD and SCZ. The PRS is then compared—using correlation values—to each individual’s score on each behavioral measure. Computation of PRS scores is described in Chapter 2.
Results

Each behavior is well-associated with its respective factor, and the factors are relatively independent.

If Zheng and colleagues’ proposed factors—PIMB and SIA—truly exist, PIMB and SIA should theoretically indicate groupings of items that are highly similar within each factor but dissimilar from each other. More specifically, we would expect that the scores on all the questions within one factor—for instance, PIMB—would correlate most with each other. We would also expect that behavioral questions that are placed into different factors would correlate relatively less with each other. This is to say that if any given pair of questions measure the same behavior, we would expect to find them in the same factor; on the contrary, if two questions measure unrelated behaviors, then we would expect to find them in different factors. If we can observe this pattern among our ten selected questions mentioned in Table 3.1, such a result would indicate that the previously described factor structure underlying these common ASD diagnostic tools are replicable and real.

Within the SSC cohort, each pair of PIMB factor questions were moderately to highly correlated (average $r(1460) = 0.46$; Figure 3.2, dark cyan triangles). This was also true for all pairs of SIA factor questions (Fisher’s Z-transform, average $r(1460) = 0.38$, Figure 3.2, dark cyan triangles). On the other hand, we observed comparatively low, but not zero, correlations across all possible PIMB-to-SIA question pairs (average $r(1460) = 0.12$, Figure 3.2, light cyan square). All pairs of correlation values and 95% CI can be found in SI Table 3.1. In conclusion, because questions within PIMB are well-correlated with each other, questions within SIA are well-correlated, but questions from PIMB are not well correlated with those of SIA, the factor structure of the questionnaire was indeed present in our cohort. Thus, PIMB and SIA are factors
that contain highly related questions within the factor and distinct questions between each factor. This result both replicates and affirms the previous findings sourced from the latent factor analysis of SRS, ADI-R, and ADOS questions.
Figure 3.2: Correlation matrix of behavioral questions, factors, and ASD/SCZ polygenic risk scores across all samples. Each gray box represents the variables of the correlation in boxes. Negative correlations are shaded with red, and positive correlations are shaded in different colors for each group of results. Cyan indicates positive correlations between behavioral questions, green indicates positive correlations between PRS scores and behavioral questions, and blue indicates positive correlations between each behavior and factor score (sum). Absence of correlation ($r = 0$) is shaded white. Intensity of colors represent $r$ values. All fit lines are least-squares and have a physical slope that equals the Pearson correlation coefficient, since axes are square and constrained to span equivalent numbers of SDs.
Factors are well-captured by categorical sums.

Next, given the distinctive factor structure of PIMB and SIA, I now simplify this information into one summary statistic. A summary statistic of each factor is able to (1) capture information about the factor as a whole, and (2) relate this value—which represents a group of specific, similar questions—to a person’s genetic information.

A good descriptor of each factor should, ideally, correlate well with each of the questions within the factor of interest but correlate little with the other factors. Concretely, a descriptor PIMB should correlate well with all five questions that make up PIMB but should not correlate well with any of the five questions that make up SIA. To create a measure that behaves as described, I first employed a factor sum—that is, the sum of all the questions within each factor—to describe the performance of any given person within PIMB and SIA. With this method, I obtain one summary statistic for PIMB and another summary statistic for SIA for every subject. Because all questions within a factor were scored on an integer scale of 0 to 3 and there were five total questions, the maximum score was 15 and minimum score was 0, with a higher score indicating higher demonstrated ability of the given factor.

A factor sum is not the only way to summarize the information across all the questions in a factor. Due to this fact, I next assessed whether a factor sum would be sufficient for assessing factor relationships using correlational analysis with every behavioral question. In correlating the individual factor sums (“PIMB sum” and “SIA sum”) and every question utilized in our analysis, the PIMB factor sum correlated well with each of the five questions in PIMB (average $r(1460) = 0.46$), while it correlated relatively little with the five questions in SIA (average $r(1460) = 0.46$), Figure 3.2, blue boxes, left column. Thus, PIMB factor sum mirrors our expectations of what a good descriptor of PIMB should be. Similarly, SIA sum correlated well with each of the five questions in SIA (average $r(1460) = 0.46$), but did not correlate as much with questions in the
PIMB factor (average $r(1460) = 0.46$) **Figure 3.2**, blue boxes, right column. This again replicates the factor relationship that we expect to observe, and it affirms the summary value's utility as a sufficient descriptor.

Now having established that factor sums are sufficient summaries of PIMB and SIA, I employ these summary statistics to test the relationship between PIMB, SIA, and gene sets in the form of PRS.

**ASD PRS is not associated with PIMB or SIA in the SSC sample.**

Using factor sums of PIMB and SIA, we assessed both the correlation coefficient with a 95% CI and the proportion of variance in behavior explained by the ASD PRS ($R^2$). Surprisingly, there was no relationship between ASD PRS and either PIMB ($r(1460) = -.02, 95\%\ CI [-.07, .04], p = .0545; R^2 \approx 0$) or SIA ($r(1460) = .01, 95\%\ CI [-.04, .06], p = .0730; R^2 \approx 0$; **Figure 3.2**, green-red boxes, upper row). This result is unexpected, as it suggests that—among other possibilities—there is no clear relationship between biological liability to ASD and modification of behavior (represented by PIMB) or social initiation (represented by SIA) in our sample. When we observe the relationships between ASD PRS and all the questions individually, we observe a slightly negative correlation with Q37, ($r(1460) = -.06, 95\%\ CI [-.11, -.01], p = .0170$), but there is not enough power to detect it after applying the Bonferroni multiple hypothesis testing correction (Bonferroni correction, $p > .0025$).

**SCZ PRS is negatively associated with PIMB, but not SIA.**

Unlike ASD, SCZ PRS shows a slightly negative relationship between genetic risk for SCZ and the PIMB factor sum ($r(1460) = -.07, 95\%\ CI [-.12, -.02], p = .0048$; **Figure 3.2**, top right). Unpacking this relationship further, we observe a small but consistent effect of SCZ risk
with all of the top three questions within the PIMB factor ($r(1460) = -.09, 95\% \text{ CI} [-.14, -.04]$; $r(1460) = -.08, 95\% \text{ CI} [-.13, -.03]$; $r(1460) = -.06, 95\% \text{ CI} [-.11, -.01]$; in order of Figure 3.2, green-red boxes, bottom row), which suggests a small but robust relationship between SCZ PRS and the PIMB factor. That is, SCZ PRS is modestly predictive of PIMB in an ASD-ascertained sample, but not SIA. Interestingly, there also exists a slight positive relationship between SCZ PRS and the SIA factor item “Imaginative play with peers” ($r(1460) = .05, 95\% \text{ CI} [.00, .10], p = .0627$), although this effect is small and does not pass the significance test ($p > .05$). If this effect does exist but our sample is underpowered to detect this effect after multiple hypothesis correction, this result would be consistent with typical SCZ symptoms of imaginative hallucinations.

Overall, because higher SCZ PRS is associated with worse performance in the PIMB factor—especially the individual questions “Has difficulty relating to peers,” “Is socially awkward,” and “Has difficulty making friends”—but not SIA, a scientific consequence is that PIMB and SIA may indeed measure an underlying aspect of social behavior that is characteristically different in some way. In this case, as SCZ PRS is associated with PIMB but not SIA, genetic factors that relate to SCZ are also related to peer interaction, but not social initiation. This result warrants future investigation of what PIMB and SIA are differentially measuring in schizophrenia, which will be further discussed later in this paper. One possible explanation may be that SIA and PIMB are associated with another variable, such as case IQ, that is also associated with SCZ PRS.

These results, together, establish the relationship between PRS and social behavior in the general SSC sample. Yet, another important question remains: how does ancestry affect our results? Genomic research data have, to this point, included mostly Caucasian (or, in rarer cases, East Asian) individuals; therefore, PRS is not likely to generalize to individuals of all
ancestries. Despite this critical limitation of PRS, the exact extent to which the lack of
generalizability affects experimental outcomes on social behavior across different ancestries
has not been extensively studied. In order to address this knowledge gap, I next decrease the
sample set to only individuals from a European or non-European ancestry in order to compare
how different ancestries affect the scientific conclusions one might draw.

**PRS scores in non-European samples tend to be larger and highly varied.**

Much of genetic research intentionally excludes non-European subjects in order to study
a more genetically homogenous population. By observing non-European populations and
European populations separately, one can observe both (1) the typically scrutinized, highly
homogenous population (subjects of European ancestry only) in addition to (2) the pervasively
understudied, heavily marginalized population (subjects of non-European ancestries). This
approach enables us to observe inherent ethnic biases of the behavioral and genetic measures
commonly employed in behavioral-genetic research. This resulted in two subsets of European
(n = 1,191) and non-European (n = 271) samples.

Theoretically, all individuals who are clinically diagnosed with the same disorder with
similar ASD severity (verbal with fluent speech) should have similarly distributed PRS,
especially since PRS aims to summarize one’s *overall risk* for having ASD based on the
aggregate effects of genome-wide common variation that a person has. However, this was not
the case. First, upon comparing the PRS distribution of both ASD and SCZ in European and
non-European samples, we consistently observed far greater heterogeneity in the
non-European sample relative to the European sample (Figure 3.3). Second, among our
ASD-ascertained samples, the European sample maintained—on average—a lower overall
PRS for both ASD and SCZ in comparison to non-European samples. These differences in PRS
distribution across the two populations reflect an inherent bias in PRS: while those diagnosed with ASD from a European ancestry have relatively similar, normally distributed PRS, those with ASD who are of a non-European ancestry have a relatively heterogeneous PRS due to the group’s varied ancestries.

![Figure 3.3: Distribution of ASD (top) and SCZ (bottom) polygenic risk scores for European and non-European samples.](image)

Compared to European (red) samples, non-European samples (blue)—on average—maintain a higher mean PRS with greater variability.

**Factor structure is less defined in the non-European subset.**

The aforementioned difference in PRS in European and non-European populations might have been expected, as it is known that PRS is most suitable for populations that mimic
the samples that were studied in order to create it. Does the same hold for behavioral tests?

Our results show that, indeed, the factors PIMB and SIA are less well-defined in non-European samples compared to the larger European sample (Figure 3.4 C, D). A well-defined factor, as observed in the European sample, would show higher correlations within the factor (Figure 3.4 A, B, dark triangles) and lower correlation across the two factors (Figure 3.4 A, B, light square). However, this pattern is much less defined in non-European samples (Figure 3.4 C, D). Instead of the expected factor structure, non-European samples’ scores on behavioral questions show small to moderate relationships across all questions across both factors. The only outlier to this trend is Q13, “Is awkward in turn-taking,” which was associated with questions in PIMB but not SIA. This finding suggests that the question-by-question factor analysis results by Zheng et al. is not fully generalizable to a non-European population.

Interestingly, PIMB and SIA factor sums are still able to distinguish between PIMB and SIA relatively well in non-European samples even despite unclear factor structure in comparison to the European samples. This vagary in the factor structure is reflected somewhat by the fact that the factor sums do not as clearly distinguish between PIMB and SIA as do European samples’ factor sums (Figure 3.4, right columns). Thus, it may be reasonable to suppose that while the individual questions in a non-European sample are not as clearly related to PIMB or SIA, the factors themselves are robust enough to capture something distinctively meaningful about behavior in both European and non-European populations.
Figure 3.4: Correlation matrices of PRS and social behavior for European (top) and non-European (bottom) samples. (A), (C): Correlation matrices with all data points visible. Negative correlations are shaded with red, and positive correlations are shaded in different colors.
for each group of results. Cyan indicates positive correlations between behavioral questions, green indicates positive correlations between PRS scores and behavioral questions, and blue indicates positive correlations between each behavior and factor score (sum). Absence of correlation \( (r = 0) \) is shaded white. Intensity of colors represent \( r \) values. (B), (D): Simplified versions of the correlation matrix on the left with correlational values (upper right triangle) and 95% CI (bottom left triangle) visible. Non-significant correlations at the \( p = .05 \) threshold are also shaded white. Note that behavioral results for European and non-European samples have approximately equal variance, but PRS do not share equal variance; in addition, the axes range for panels (A) and (C) are different, with the ranges determined by the centrality and range of given data points, but are not labelled due to resolution. All fit lines are least-squares and have a physical slope that equals the Pearson correlation coefficient, since axes are square and constrained to span equivalent numbers of SDs.

**Associations between SCZ PRS and social behavior are driven by the non-European sample.**

Given that factor sums are relatively good at capturing either PIMB or SIA performance across individuals of diverse ancestries, how do these factor sums fare in establishing a relationship between genes and behavior? Ancestry-specific analysis suggests that some of the negative relationship between SCZ PRS and individual questions within PIMB are driven by the non-European population. This effect was most pronounced in Q37 “Has difficulty relating to peers” \( (r(269) = -.14, 95\% \text{ CI } [-.25, -.02], p = .0239) \) and Q33 “Is socially awkward” \( (r(269) = -.12, 95\% \text{ CI } [-.24, -.01], p = .0408) \). The associations between SCZ PRS and these same two questions did not exist to the same extent in the European subset \( (r(1,189) = -.05, 95\% \text{ CI } [-.11, .00], p = .0692; r(1,189) = -.06, 95\% \text{ CI } [-.11, .00], p = .0523, \) for Q37 and Q33 respectively). While other small differences between the two sample subsets and their respective significant correlations exist, these differences are small and more likely attributed to differences in sample size. Furthermore, these associations in the non-European population, although often statistically insignificant, led to a difference of almost 2% of behavior explained by these genes.
versus 0.3%. Together, these results indicate that this small subset of the population drives the effect between SCZ PRS and PIMB factor, with a much larger portion of the variation in behavior explained by the PRS.

From these results, a few questions arise: what is causing this difference in the relationship between SCZ PRS and its association with behavior? Is this a result of ancestry-specific genetic effects, or is it measuring a structural bias within genomic research? To answer these questions, we turn to known facts about the relationship of ASD and SCZ: they share many overlapping risk genes, and are both highly heritable, which suggests similar underlying biological architectures. One might expect, then, that PRS for ASD and SCZ would be highly correlated. Any differences in the degree of correlation between ASD and SCZ PRS between the two populations may suggest inherent biases in PRS measurements that may consistently inflate or deflate predicted risk in non-European populations.

The association between ASD and SCZ PRS is greater in non-European samples.

Based on the fact that ASD and SCZ PRS are expected to be correlated, I assessed how genetic risk factors for ASD and SCZ relate differentially among people from different ancestries. First, a positive correlation is expected between ASD and SCZ PRS scores for both ancestries. This was indeed the case for both European \( r(1,189) = .23, \text{95\% CI [.18, .28]} \) and non-European \( r(269) = .63, \text{95\% CI [.54, .69]} \) samples. However, it is evident that ASD and SCZ PRS are much more predictive of each other in the non-European sample, with only 5.3% of variation in genetic risk for SCZ being explained by ASD risk in European samples, but 39.7% of variation in genetic risk being accounted for by ASD risk in non-European samples (Fisher's Z transformation and two-sample z-test, \( p < .0001 \)). Such an extreme difference in SCZ risk predictivity by ASD PRS across European and non-European populations is surprising,
especially given that both populations’ PRS values are computed in the same manner for all subjects. Together, these results suggest that social behavior scores, ASD PRS, and SCZ PRS all behave differently in non-European populations compared to European populations.

Discussion and Future Direction

In this analysis, I investigated the relationship between ASD PRS and SCZ PRS on a variety of behavioral questions and their associated factors, PIMB and SIA, which each describe the child’s interaction with peers or prosociality, respectively. First, overall, ASD PRS was not associated with either PIMB or SIA, but SCZ PRS was slightly associated with PIMB. Next, I discovered that while most assumptions, including stratification of factors, high utility of factor sums, and association between ASD PRS and SCZ PRS are indeed observed, the extent to which these assumptions hold are different between samples of European ancestry and samples of non-European ancestry. Furthermore, PRS scores are far more varied—and typically higher—in non-European samples. Additionally, despite the non-European sample’s smaller size, we showed that this non-European sample contributed more to the overall association between PRS and PIMB factor sum in the full-sample analysis when compared to the contributions of the European sample alone. This suggests that when we assess the bias in ancestry—or race—within genetic and behavioral research, we find non-convergent or heterogeneous results in the two populations.

These results bring up more questions than what they answer. First, if PRS scores correlate with ASD and SCZ diagnosis, and if ASD and SCZ are diagnosed by behavioral symptoms, why does ASD PRS not correlate with ASD-related behavior, and why does SCZ PRS only correlate slightly with SCZ-related behavior? One possible explanation is that the SSC
sample has been ascertained for ASD. Thus, the degree of variation in behavior among this cohort is very limited, which limits the signal between genes and behavior that we can observe. This, in turn, may explain why there was a relationship between SCZ PRS and behavior, but not ASD PRS and behavior—this sample was not ascertained for SCZ, so the SCZ PRS was not constrained, thereby enabling the detection of a relationship between SCZ PRS and behavior. This suggests that PRS-based approaches in investigating genotype-phenotype relationships are best utilized for the general population, not a niche group of individuals who are all diagnosed with the same condition. However, this analysis may prove useful for elucidating the genetic of dual diagnoses: by studying SCZ PRS and its relationship to ASD behaviors in an ASD-ascertained sample, one might be able to mediate between the compounded, potentially nonlinear effects of SCZ on an individual with an ASD and SCZ dual diagnosis.

Another possibility for the lack of observed relationship between ASD PRS and behavior is that, for those who have enough genetic liability to be diagnosed with ASD, the individual variation in social ability at that point is mostly mediated by environment, not genetic risk. This may explain why ASD genes have no effect on ASD-related behaviors within an ASD-ascertained population. This result, therefore, does not apply to how ASD PRS may relate to behavior in the broader population. A third possibility, which rests on the goodness of our measures, is that PRS and/or behavioral measures are not precise, reliable, or valid enough to detect associations to the extent that noise overwhelms the signal. This possibility is less likely than the first two given the relative homogeneity of scores in the European population, who were also the majority of the samples in this study; however, considering the relatively high variation in the non-European samples, this possibility should not be ignored. In fact, this latter possibility is more likely to hold in a similar study conducted entirely on a non-European population. It is also important to note that the validity of PRS—as it pertains to
psychometrics—has not been extensively tested, raising questions about what PRS truly measures (Janssens, 2019).

Third, why are the PRS for European samples less varied and smaller on average compared to PRS for non-European samples? Similarly, why are ASD PRS and SCZ PRS more correlated for a non-European population? While the larger variation of PRS in non-European samples might be directly caused by the greater diversity of this very subset, the higher average PRS in comparison to the European subset is somewhat concerning from a scientific standpoint. A higher mean PRS indicates that, on average, these individuals harbor more genetic risk than do European samples. Why are non-European individuals more likely to be assigned a higher risk score for a disorder? Does this fact indicate something inherent about PRS computation methods, like its lack of sensitivity to LD in non-European populations? Or, does this fact indicate that sampling methods for non-European individuals are flawed such that individuals with greater genetic risk for ASD are more likely to be recruited? This puzzling result should be further investigated in the future, as it has critical implications on—as one futuristic example—prediction algorithms using PRS of whether a person might have a severe psychiatric condition.

It is important to note some analysis-related considerations and contextualize our findings relative to other literature. First, some differences exist in the behavioral data presented here compared to the factor analysis by Zheng et al. In this study, rather than simplifying all behavioral scores to 0, 1, and 2 by collapsing scores of both 2 and 3 into “2”, we kept all scores between 0 and 3. This practice, which is unconventional, may have introduced factor-related differences within our interpretation, although it expanded the range of behavioral values in correlational analysis. Second, with regard to genetic data, it is common practice to control for the principal components of each subject’s genome in order to mitigate ancestry-related and
miscellaneous genome-related effects. For instance, the pattern of inheritance of genes is not perfectly random—there are genes that tend to be linked more frequently together (LD), ancestry-specific effects, and more. Thus, by extracting the principal component of a person’s genomic information and controlling for it, we can get rid of the effects of various artifacts including population stratification, ancestry, and linkage disequilibrium (Bhattacharjee et al., 2010). This would be an important next step to increase the accuracy of this study.

Conclusion

Genotype-phenotype investigations using PRS and a variety of behavioral assessments often lead to eye-opening findings. While these genetic and behavioral investigative approaches can pave the way for a future filled with innovations in modern medicine, including precision medicine in psychiatry, improvements in the diagnostic criteria, paradigmatic shifts in our conception of psychopathology, or a new way to understand human variation, what we measure is ultimately limited by for whom these measures were developed. Until our measures and methodologies are better suited for a diverse population, applying our current Eurocentric translational research directly to medical practice for the general population may be unethical, as it is sure to perpetuate well-known health disparities. This serves as a wake-up call in scientific communities to create measures, both psychological and biological, that serve the good of the whole population and to employ research studies beyond the norm of homogenizing sample sets—whether it be facial stimuli for behavioral tests or limiting analysis to European samples in genomic research—under the guise of science.
Literature Cited


Cross-Disorder Group of the Psychiatric Genomics Consortium. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics, 45*(9), 984–994. https://doi.org/10.1038/ng.2711


https://doi.org/10.1016/j.biopsych.2017.08.017


Kynast, J., Quinque, E. M., Polyakova, M., Luck, T., Riedel-Heller, S. G., Baron-Cohen, S., Hinz,


McIntosh, A. M., Job, D. E., Moorhead, W. J., Harrison, L. K., Whalley, H. C., Johnstone, E. C.,


https://doi.org/10.1016/j.cell.2019.12.036


https://doi.org/10.1001/jamanetworkopen.2019.21644


Zhang, J.-P., Robinson, D., Yu, J., Gallego, J., Fleischhacker, W. W., Kahn, R. S.,

Author Contributions

For Chapter 1, H.K., L.G., J.B.W., R.S. designed and performed study; J.K., J.B.W., and L.G. developed MRMET; S.C. created the stimuli used for MRMET development; H.K., R.S., and J.B.W. analyzed data; H.K., R.S., J.K., and J.B.W. created figures; H.K., J.K., L.G., S.C., R.S., and J.B.W wrote the paper.

H.K. wrote Chapter 2.

For Chapter 3, H.K., designed the study, analyzed data, created figures, and wrote the paper.
Supplemental Information

SI Table 1.1: Word list for RMET with gender stereotype indicators for each image.

<table>
<thead>
<tr>
<th>RMET</th>
<th>Items</th>
<th>Answer choices</th>
<th>Demographics</th>
<th>Gender Stereotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Correct Race</td>
<td>(approximated)</td>
</tr>
<tr>
<td>practice</td>
<td>jealous</td>
<td>panic</td>
<td>arrogant</td>
<td>hateful</td>
</tr>
<tr>
<td>2</td>
<td>playful</td>
<td>comforting</td>
<td>irritated</td>
<td>bored</td>
</tr>
<tr>
<td>3</td>
<td>terrified</td>
<td>upset</td>
<td>arrogant</td>
<td>annoyed</td>
</tr>
<tr>
<td>4</td>
<td>joking</td>
<td>flustered</td>
<td>desire</td>
<td>convinced</td>
</tr>
<tr>
<td>5</td>
<td>joking</td>
<td>insisting</td>
<td>amused</td>
<td>relaxed</td>
</tr>
<tr>
<td>6</td>
<td>aghast</td>
<td>sarcastic</td>
<td>worried</td>
<td>friendly</td>
</tr>
<tr>
<td>7</td>
<td>apologetic</td>
<td>fantasizing</td>
<td>impatient</td>
<td>alarmed</td>
</tr>
<tr>
<td>8</td>
<td>despondent</td>
<td>comfortable</td>
<td>uneasy</td>
<td>dispirited</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td>relieved</td>
<td>shy</td>
</tr>
<tr>
<td>10</td>
<td>annoyed</td>
<td>hostile</td>
<td>horrified</td>
<td>preoccupied</td>
</tr>
<tr>
<td>11</td>
<td>cautious</td>
<td>insisting</td>
<td>bored</td>
<td>aghast</td>
</tr>
<tr>
<td>12</td>
<td>terrified</td>
<td>amused</td>
<td>regretful</td>
<td>flirtatious</td>
</tr>
<tr>
<td>13</td>
<td>indifferent</td>
<td>embarrassed</td>
<td>sceptical</td>
<td>dispirited</td>
</tr>
<tr>
<td>14</td>
<td>decisive</td>
<td>anticipating</td>
<td>threatening</td>
<td>shy</td>
</tr>
<tr>
<td>15</td>
<td>irritated</td>
<td>disappointed</td>
<td>depressed</td>
<td>accusing</td>
</tr>
<tr>
<td>16</td>
<td>contemplative</td>
<td>flustered</td>
<td>encouraging</td>
<td>amused</td>
</tr>
<tr>
<td>17</td>
<td>irritated</td>
<td>thoughtful</td>
<td>encouraging</td>
<td>sympathetic</td>
</tr>
<tr>
<td>18</td>
<td>doubtful</td>
<td>affectionate</td>
<td>playful</td>
<td>aghast</td>
</tr>
<tr>
<td>19</td>
<td>decisive</td>
<td>amused</td>
<td>aghast</td>
<td>bored</td>
</tr>
<tr>
<td>20</td>
<td>arrogant</td>
<td>grateful</td>
<td>sarcastic</td>
<td>tentative</td>
</tr>
<tr>
<td>21</td>
<td>dominant</td>
<td>friendly</td>
<td>guilty</td>
<td>horrified</td>
</tr>
<tr>
<td>Items</td>
<td>Answer choices</td>
<td>Demographic information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>----------------</td>
<td>------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SI Table 1.2: Word list for MRMET with demographic indicators where available. * = age approximated from education information</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MRMET</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Trial</strong></td>
<td><strong>Randomized trial order</strong></td>
<td><strong>Option 1</strong></td>
<td><strong>Option 2</strong></td>
<td><strong>Option 3</strong></td>
</tr>
<tr>
<td>practice</td>
<td>anxious</td>
<td>disappointed</td>
<td>shocked</td>
<td>concerned</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>nervous</td>
<td>sarcastic</td>
<td>curious</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>threatening</td>
<td>disappointed</td>
<td>panicked</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>indifferent</td>
<td><strong>relieved</strong></td>
<td>puzzled</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>excited</td>
<td>embarrassed</td>
<td>interested</td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>threatening</td>
<td>panicked</td>
<td>concerned</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>5</td>
<td>confident</td>
<td><strong>confused</strong></td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>curious</td>
<td>friendly</td>
<td>sarcastic</td>
</tr>
<tr>
<td>8</td>
<td>31</td>
<td>amused</td>
<td>excited</td>
<td>curious</td>
</tr>
<tr>
<td>9</td>
<td>36</td>
<td>irritated</td>
<td>threatening</td>
<td>concerned</td>
</tr>
<tr>
<td>10</td>
<td>13</td>
<td>interested</td>
<td>sarcastic</td>
<td>embarrassed</td>
</tr>
<tr>
<td>11</td>
<td>23</td>
<td>indifferent</td>
<td>preoccupied</td>
<td>uneasy</td>
</tr>
<tr>
<td>12</td>
<td>14</td>
<td>friendly</td>
<td>sarcastic</td>
<td>curious</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>sarcastic</td>
<td>interested</td>
<td>embarrassed</td>
</tr>
<tr>
<td>14</td>
<td>30</td>
<td>amused</td>
<td>excited</td>
<td>curious</td>
</tr>
<tr>
<td>15</td>
<td>34</td>
<td>relieved</td>
<td>indifferent</td>
<td>puzzled</td>
</tr>
<tr>
<td>16</td>
<td>25</td>
<td>panicked</td>
<td>threatening</td>
<td>disappointed</td>
</tr>
<tr>
<td>17</td>
<td>6</td>
<td>confused</td>
<td>hostile</td>
<td>threatening</td>
</tr>
<tr>
<td>18</td>
<td>24</td>
<td>thoughtful</td>
<td>excited</td>
<td>fantasizing</td>
</tr>
<tr>
<td>19</td>
<td>33</td>
<td>curious</td>
<td><strong>amused</strong></td>
<td>excited</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td>confused</td>
<td>hostile</td>
<td>threatening</td>
</tr>
<tr>
<td>21</td>
<td>19</td>
<td>uneasy</td>
<td>puzzled</td>
<td>indifferent</td>
</tr>
<tr>
<td>22</td>
<td>20</td>
<td>indifferent</td>
<td>preoccupied</td>
<td>puzzled</td>
</tr>
<tr>
<td>23</td>
<td>4</td>
<td>uneasy</td>
<td>confused</td>
<td>confident</td>
</tr>
<tr>
<td>24</td>
<td>2</td>
<td>nervous</td>
<td>ashamed</td>
<td>affectionate</td>
</tr>
<tr>
<td>25</td>
<td>32</td>
<td>excited</td>
<td>disappointed</td>
<td>curious</td>
</tr>
<tr>
<td>26</td>
<td>8</td>
<td>accusing</td>
<td>confused</td>
<td>cautious</td>
</tr>
<tr>
<td>27</td>
<td>29</td>
<td>d</td>
<td>excited</td>
<td><strong>amused</strong></td>
</tr>
<tr>
<td>28</td>
<td>15</td>
<td>friendly</td>
<td>sarcastic</td>
<td>curious</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td>affectionate</td>
<td>ashamed</td>
<td>friendly</td>
</tr>
<tr>
<td>30</td>
<td>9</td>
<td>curious</td>
<td>distrustful</td>
<td>hostile</td>
</tr>
<tr>
<td>31</td>
<td>3</td>
<td>concerned</td>
<td>confused</td>
<td>defiant</td>
</tr>
<tr>
<td>32</td>
<td>28</td>
<td>depressed</td>
<td>preoccupied</td>
<td>confused</td>
</tr>
<tr>
<td>33</td>
<td>11</td>
<td>excited</td>
<td>embarrassed</td>
<td>interested</td>
</tr>
<tr>
<td>34</td>
<td>21</td>
<td>preoccupied</td>
<td>indifferent</td>
<td>uneasy</td>
</tr>
<tr>
<td>35</td>
<td>37</td>
<td>enraged</td>
<td>angry</td>
<td>threatening</td>
</tr>
<tr>
<td>36</td>
<td>22</td>
<td>indifferent</td>
<td>preoccupied</td>
<td>puzzled</td>
</tr>
<tr>
<td>37</td>
<td>18</td>
<td>hostile</td>
<td>threatening</td>
<td>shocked</td>
</tr>
</tbody>
</table>
SI Table 3.1: Correlation, p-value, and 95% CI for all analyzed correlations (for European-only, non-European-only, and full SSC sample) can be found in:
https://drive.google.com/drive/folders/1DeKjTk3V33sHqWIVjD4dsZfkm7Ekp6RT?usp=sharing

SI Figure 3.1: Visualization of the correlations (upper right triangle), and 95% CI (bottom left triangle) between ASD PRS, SCZ PRS, and all behavioral questions/factors. This is the same figure as Figure 3.2, but visualized such that individual correlation values are visible. Color indicates correlation values. Non-significant correlations and r = 0 are left white.
Appendix

Genomic Modelling

In Chapter 3, there were notably small effects by genes on behavior. This is indeed the case for many risk loci detected by GWAS and other statistical analyses that aim to see the effects of individual genes or groups of genes on a particular phenotype. In order to improve our ability to detect small differences — that is, increase power — other statistical paradigms might be employed to study these relationships. For instance, multivariate methods in statistical genomics that utilize phenotypic data and other a priori biological information to improve power is one approach that may assist with decreasing the amount of genetic effect that heritability does not account for, which will be further explored in this Appendix. The aim of this Appendix is to provide a set of tools that may be of interest for future implementation when continuing genotype-phenotype analyses on gene sets that have small effects on behavior or phenotypes. These are alternative methods that could be implemented in a future study.
Regression Models

**MARV (Multi-phenotype Analysis of Rare Variants)**

MARV is a reverse regression method that, by definition, reverses the typical prediction of phenotype from genotype. Just as linear regression enables utilization of multiple genotypes to predict one phenotype, reverse regression enables prediction of genotype from multiple phenotypes simultaneously. Unlike other reverse regression models, however, MARV focuses on mutational load of risk alleles calculated by proportion of rare variants that carry minor alleles—known as burden—at each rare variant loci. MARV focuses on one genomic region at a time, each of which is pre-specified with gene boundaries and cutoff regions. For each region, the proportion of minor alleles at rare variants within the region is computed; this value serves as the predicted output. The minor allele to rare variant proportion is then modelled as a linear combination of $K$ phenotypes, thereby allowing the $K$ phenotypes to serve as the predictors. In other words, MARV is particularly suited for analysis of rare variants and their effects on multiple phenotypes. Upon aggregating this data across all subjects, linear regression is employed using the provided phenotypes as predictors for the model. This model is fine-tuned by estimating the likelihood of contribution of each individual to the outcome as a function of the number of successfully genotyped or imputed rare variants. In other words, MARV weighs each individual by the number of rare variants that have been genotyped or imputed within the region of interest. Finally, MARV interchanges various predictor combinations to assess which model yields the best Bayesian Information Criterion (BIC).

Mathematically, this model can be portrayed as $\frac{r_i}{n_i} = \alpha + \beta y_i + \epsilon_i$, with $r_i/n_i$ indicating the proportion of minor alleles for individual $i$, $y_i$ indicating vector of phenotypes for individual $i$, $\beta$ denoting matrix of regression coefficients, $\epsilon_i$ representing a value derived from the
multivariate normal distribution from \((0, \sigma^2)\), where \(\sigma^2\) indicates the covariance matrix. Following this model, the weighted likelihood ratios of the fitted model is computed and significance testing is performed against the null hypothesis that \(\beta_1, \beta_2, \ldots, \beta_k = 0\).

Advantages of MARV include its ability to incorporate binary and continuous phenotypes while enabling direct interpretation of affected phenotypes, as well as its improved power relative to univariate burden tests. However, its incorporation of a burden test, which is useful for identifying rare variants when large numbers of such variants are causal with the same directionality of effect, limits its power in cases in which such conditions are not met.

**SCOPA/META-SCOPA (Software for COrelated Phenotype Analysis)**

SCOPA is another reverse regression model that incorporates multiple correlated phenotypes to predict SNPs. In contrast to MARV, however, SCOPA focuses on each SNP present rather than mutational burden within a region of interest; thus, SCOPA relies on typed or imputed SNPs in an additive model. SCOPA then employs model selection based on BIC, as does MARV, to choose the best model that describes the relationship between all phenotypes and each SNP. In other words, for individual \(i\) with phenotype vectors \(k_i\) through \(k_j\), SCOPA predicts their genotype \(G_i\) at a SNP. Thus, the model can be summarized as

\[
G_i = \alpha + \sum_j \beta_j y_{ij} + \epsilon_i
\]

as an additive model for the number of minor alleles. SCOPA sets \(\beta_j\) as the effect of the \(j\)th phenotype on genotype at the SNP, and \(\epsilon_i \sim N(0, \sigma^2)\), where \(\sigma^2\) is the residual variance. Notably, this model is essentially the equivalent to MARV with the minor difference that the prediction involves each SNP rather than the proportion of minor alleles.

SCOPA is unique in its providence of a fixed-effects meta-analysis tool, called META-SCOPA. META-SCOPA utilizes resultant SCOPA association summary statistics in the
context of aggregated summary statistics across GWAS. When a number of GWAS with the same set of correlated phenotypes are provided, META-SCOPA computes the maximum likelihood estimates of phenotypic effect for each SNP. The metaanalysis thus takes the form of synthesis of regression slope, as well as BIC for each model across GWAS to estimate error.

Limitations of SCOPA are similar to that of MARV and other reverse regression models—interpretation of the model is not straightforward, as the model does not assess the effect of each SNP on each phenotype.
PHARAOH-GEE

PHARAOH-GEE is a multivariate extension of the univariate PHARAOH—a method for assessing the association of all provided pathways with each given phenotype—with the aid of generalized estimating equations (GEE). While PHARAOH assesses the relationship between various pathways and one given phenotype, PHARAOH-GEE assesses the relationship between various pathways and multiple correlated phenotypes in a single model. Furthermore, PHARAOH-GEE takes into account the relatedness of phenotypic variables with the help of GEE. PHARAOH-based methods are well-suited for detecting the effects of rare variants because [xx]. [other characteristics of PHARAOH-GEE].

The computational approach of PHARAOH-GEE is as follows: first, the genetic variants are scored through a minor allele frequency (MAF)-based method, such as the inverse minor allele frequency: $\omega = \sqrt{\text{MAF} (1 - \text{MAF})}$. Next, genes are scored by taking the sum of the scores of all overlapping genetic variants. Gene-sets are modeled as a function of their constituent genes, with each gene-set is considered as an independent pathway. Lastly, in the parameter estimation step, a vector of gene weights $w$ and pathway weights $\beta$ are fit to the data with regularization. The determined parameters are then compared against the null distribution, which is computed through random permutations of [xxx]. The null hypothesis for PHARAOH-GEE states that the $k$th pathway has no detectable association with any of the $Q$ phenotypes ($H_0: \beta_{k1} = \beta_{k2} = \ldots = \beta_{kQ} = 0$). Following model assumptions, PHARAOH-GEE is limited in analysis to binary, count and exponentially distributed phenotypes.

PHARAOH-GEE extends the methods of PHARAOH by incorporating all pathways into a single model rather than testing them independently as most gene-set analyses do. Since the
model only produces one set of parameters, only one significance test must be performed, removing the need for multiple testing corrections. For instance, the univariate PHARAOH is limited to detecting correlations between pathways and adjusting their combined association to the single phenotypic variable, whereas PHARAOH-GEE is able to associate all pathways in relation to all phenotypes at once.

Generalized estimating equations (GEE) in PHARAOH-GEE serves to take into account pre-existing associations across multiple phenotypes. Because of this characteristic, the use of GEE in the analysis of biological data has been well established. Basic linear regression (as in LCT/NLCT) naively assumes independence between phenotypes, potentially inflating the influence of highly correlated phenotype measurements. On the contrary, GEE resolves this issue by modelling the regression between a measurement and an outcome separately from the correlations between different outcomes. PHARAOH-GEE thereby assumes a different correlational structure than LCT or NLCT and is more suited to the analysis of phenotypes that are known to be correlated with one another, such as clustered phenotypes, repeated measurements, and phenotypes that assess the same underlying biological construct. An added benefit to the use of GEE is that its results are guaranteed to be consistent regardless of how closely the data follows the expected correlational structure, although violations of the model assumptions will negatively affect power. Overall, PHARAOH-GEE accounts for the effects of the relationships between pathways as well as the relationships between phenotypes and holds promise for determining rare variants for multiple correlated phenotypes.

In sum, these analytic approaches to genotype-phenotype relationships enable discovery of risk loci, and reversely, the prediction of genotype based on phenotypic characteristics.