Title: Enhancing the precision of our understanding about mentalizing in adults with autism

Authors: Michael V. Lombardo[1,2,3]*, Meng-Chuan Lai[3,4,5]*, Bonnie Auyeung[3,6], Rosemary J. Holt[3], Carrie Allison[3], Paula Smith[3], Bhismadev Chakrabarti[3,7], Amber N. V. Ruigrok[3], John Suckling[8], Edward T. Bullmore[8], MRC AIMS Consortium**, Christine Ecker[9], Michael C. Craig[9,10], Declan G. M. Murphy[9], Francesca Happé[11], & Simon Baron-Cohen[3]

Affiliations:

[1] Department of Psychology, University of Cyprus, Nicosia, Cyprus

[2] Center for Applied Neuroscience, University of Cyprus, Nicosia, Cyprus

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

[4] Child and Youth Mental Health Collaborative at The Centre for Addiction and Mental Health and The Hospital for Sick Children, and Department of Psychiatry, University of Toronto, Toronto, Canada

[5] Department of Psychiatry, National Taiwan University Hospital and College of Medicine, Taipei, Taiwan

[6] School of Philosophy, Psychology and Language Sciences, Department of Psychology, University of Edinburgh, Edinburgh, UK

[7] Centre for Integrative Neuroscience and Neurodynamics, School of Psychology & Clinical Language Sciences, University of Reading, Reading, UK

[8] Brain Mapping Unit, Department of Psychiatry, University of Cambridge, Cambridge, UK

[9] Sackler Institute for Translational Neurodevelopment, Department of Forensic and Neurodevelopmental Sciences, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

[10] National Autism Unit, Bethlem Royal Hospital, SLAM NHS Foundation Trust, UK

[11] MRC Social, Genetic, and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK

* Equal contributions

** The Medical Research Council Autism Imaging Multicentre Study Consortium (MRC AIMS Consortium) is a UK collaboration between the Institute of Psychiatry (IoP) at King's College, London, the Autism Research Centre, University of Cambridge, and the Autism Research Group, University of Oxford. The Consortium members are in alphabetical order: Anthony J. Bailey (Oxford), Simon Baron-Cohen (Cambridge), Patrick F. Bolton (IoP), Edward T. Bullmore (Cambridge), Sarah Carrington (Oxford), Marco Catani (IoP), Bhismadev Chakrabarti (Cambridge), Michael C. Craig (IoP), Eileen M. Daly (IoP), Sean C. L. Deoni (IoP), Christine Ecker (IoP), Francesca Happé (IoP), Julian Henty (Cambridge), Peter Jezzard (Oxford), Patrick Johnston (IoP), Derek K. Jones (IoP), Meng-Chuan Lai (Cambridge), Michael V. Lombardo (Cambridge), Anya Madden (IoP), Diane Mullins (IoP), Clodagh M. Murphy (IoP), Declan G. M. Murphy (IoP), Greg Pasco (Cambridge), Amber N. V. Ruigrok (Cambridge), Susan A. Sadek (Cambridge), Debbie Spain (IoP), Rose Stewart (Oxford), John Suckling (Cambridge), Sally J. Wheelwright (Cambridge), Steven C. Williams (IoP), and C. Ellie Wilson (IoP).

Abstract

Difficulties in mentalizing or theory of mind are common autism spectrum conditions (ASC). However, heterogeneity in mentalizing ability between individuals with ASC is considerable, particularly in adulthood. Parsing this heterogeneity to come to more precise understanding of which individuals have difficulty has important implications, particularly for individualized approaches in clinical and translational research applications. Here we utilize unsupervised hierarchical clustering to identify data-driven subgroups within ASC based on performance on an advanced mentalizing test, the 'Reading the Mind in the Eyes' Test (RMET). We find evidence for 2 discrete ASC subgroups that can be replicably identified across two large independent datasets. The first subgroup shows clear difficulty on the RMET, with an effect size difference compared to typically-developing controls (TD) of greater than 3 standard deviations. In contrast, the second subgroup shows little to no difficulty on the RMET compared to TD individuals. These ASC subgroups are not systematically different across a range of other variables including sex/gender, age, depression or anxiety symptoms, autistic traits, trait empathy, and autism symptom severity. Verbal IQ is slightly lower in the impaired ASC subgroup, but covarying for this does not change the effect of large difficulties in mentalizing in this subgroup. These insights enable a more precise understanding of mentalizing and may have important implications for future work that takes a more individualized approach to clinical assessment and treatment. Identification of these subgroups may also facilitate work examining multiple biological mechanisms underlying ASC in translational research.

For the last 30 years we have understood that there are profound early developmental difficulties in theory of mind/mentalizing (henceforth 'mentalizing') in autism spectrum conditions (henceforth ASC) [1]. This led to the mindblindness theory of ASC, which is one of the primary cognitive explanations behind the social-communicative difficulties that are hallmarks of ASC [2-4]. Since then, the literature on this topic has expanded considerably. For example, rather than lacking such ability altogether, there is evidence to support the idea that rudimentary explicit mentalizing ability enabling one to pass standard false belief test can develop in ASC, albeit at a much delayed rate over the lifespan [5]. Alongside such achievements in developing rudimentary explicit mentalizing skills comes enhanced everyday adaptive social behavior where mentalizing is a necessity [4]. Furthermore, there are distinctions between explicit versus implicit/automatic mentalizing processes, with the latter continuing to be atypical much later in life despite the individual possessing explicit abilities [6]. Mentalizing can also be integrated with known difficulties in the domain of self-referential cognition [7-14]. Underlying mindblindness, there are also distinct neural mechanisms for mentalizing that are affected in ASC (e.g., [15-22]; though also see [23]). These findings have generally been informative in furthering our understanding of mentalizing difficulties in ASC and for thinking about it as a theory behind the cardinal social-communicative difficulties.

Despite all these considerable strides forward in understanding mentalizing in ASC, some critical barriers remain in terms of the 'precision' of our understanding on the topic. Most, if not all, the evidence to date is based on statistical evidence about what differs *on-average* between an ASC group and a non-ASC comparison group. However, as we have come to learn over the history of studying ASC, heterogeneity is the rule, not the exception [24-26]. This means that it should come as no surprise that mentalizing is also affected by the heterogeneity within ASC. Many individuals will show evidence of some kind of deficit in mentalizing, while others may not show any difficulty or may simply mask the difficulty via compensatory mechanisms [6, 27]. This heterogeneity will also likely change throughout development as individuals on the spectrum acquire more competence in the domain [5]. Exemplifying these ideas about heterogeneity in the domain of mentalizing, a recent study by Byrge and colleagues examined inter-subject correlations in fMRI BOLD response during naturalistic viewing of videos displaying high levels of social awkwardness. This study found that group-differences in a standard case-control comparison were entirely driven by 5 outlier ASC individuals with idiosyncratic patterns of response [28]. These 5 individuals also showed behavioral deficits in the domain of mentalizing (i.e. reduced social comprehension, understanding of motivations, thoughts and feelings). This result directly reflects the fact that heterogeneity will always be present in typical case-control studies and that interpretation and replication of such studies may be challenging without knowing *a priori* what types of heterogeneous mixtures of ASC individuals are present in any one sample.

It is clear that we must move beyond omnibus statements about on-average statistical effects of group differences, towards a more 'precise' understanding of heterogeneity that can allow for more individualized approaches [29, 30]. A move towards more precision in understanding individual differences in mentalizing can also be of large relevance for clinical practice and translational research. For instance, such an approach would enable work to proceed toward the goals envisioned by the personalized medicine initiative (e.g., individualized treatment evaluation, more precise prognostic predictions). Furthermore, top-down approaches to parsing heterogeneity at the level of cognitive subtypes may also lead to important discoveries regarding differing biology or other types of compensatory mechanisms [25, 31].

In the current study we utilize an advanced mentalizing test (the 'Reading the Mind in the Eyes' Test; RMET) on adults with ASC to subgroup patients based on different profiles of performance on the test. The RMET is widely used as a core instrument for assessing difficulties in the social cognitive domain of understanding other minds, and its prominence has even warranted its listing as a core instrument in this domain within the NIMH RDoC (http://1.usa.gov/1Qs6MdI). The RMET is also widely utilized as a treatment outcome measure (e.g., [32-35]). Given that inconsistency exists behind findings from mentalizing-based treatments for ASC [36] and the tacit understanding that heterogeneity in treatment response is highly likely, it is critical to further our understanding of how mentalizing heterogeneity manifests in ASC, particularly on the RMET, and whether such insights can help in better trial design and/or in individualizing interpretations about who may benefit from such treatments.

Here we use RMET item-level response data as input into unsupervised hierarchical clustering to identify data-driven subgroups. This approach is novel with respect to how the RMET is traditionally analyzed as it utilizes the full spectrum of information contained across responses to all RMET items and does not rely on the traditional procedure of summing scores across items. We show that this approach can identify replicable data-driven discrete subgroups of adults with ASC. We also go further to characterize potential differences between RMET ASC subgroups on other measures such as sex/gender, verbal IQ (VIQ), autistic traits (using the Autism Spectrum Quotient, AQ), trait empathy (using the Empathy Quotient, EQ), depression and anxiety symptoms (using the Beck Depression and Anxiety Inventories, BDI and BAI), as well as gold standard autism diagnostic instruments (i.e. Autism Diagnostic Observation Schedule, ADOS; Autism Diagnostic Interview-Revised, ADI-R).

Materials and Methods

Discovery Dataset and Participant Characteristics

In this study we analyzed two large datasets that served as discovery and replication sets. The discovery dataset came from the Cambridge Autism Research Database (CARD) [37] and

consisted of 395 adults with ASC (178 males, 217 females) and 320 typically-developing controls (TD; 152 males, 168 females) within the age range of 18-74 years. The CARD data were collected online from two websites (www.autismresearchcentre.com, www.cambridgepsychology.com) during the period of 2007-2014. Once participants had logged onto either site, they consented for their data to be held in the Cambridge Autism Research Database (CARD) for research use, with ethical approval from the University of Cambridge Psychology Research Ethics Committee (reference No. Pre.2013.06).

CARD participants who self-reported a clinical autism diagnosis were asked specific information about the date of their diagnosis, where they were diagnosed, and the profession of the person who diagnosed them. The inclusion criterion for participants in the autism group was a clinical diagnosis of an autism spectrum condition (ASC) according to DSM-IV (any pervasive developmental disorder), DSM-5 (autism spectrum disorder), or ICD-10 (any pervasive developmental disorder) from a recognized specialist clinic by a psychiatrist or clinical psychologist. Such online self/parent-reported diagnoses agree well with clinical diagnoses in medical records [38]. Control group participants were included if they had no diagnoses of ASC, and no first-degree relatives with ASC. For both groups, participants were excluded if they reported a diagnosis of bipolar disorder, schizophrenia, eating disorder, obsessive-compulsive disorder, personality disorder, epilepsy, or an intersex/transsexual condition. Participants with a diagnosis of depression or anxiety were not excluded as these conditions are common in the general population and occur at high rates in adults with autism [24].

Replication Dataset and Participant Characteristics

The replication dataset consisted of participants from the MRC AIMS Consortium dataset (n=123 ASC; 85 male, 38 female; n=128 TD; 87 male, 41 female) within the age range of 18-52[39]. The study was given ethical approval by the National Research Ethics Committee, Suffolk, UK. All volunteers gave written informed consent. Participants were recruited and assessed at one of the three MRC AIMS centers: the Institute of Psychiatry, London; the Autism Research Centre, University of Cambridge; the Autism Research Group, University of Oxford. All participants were right-handed. Exclusion criteria for all participants included a history of major psychiatric disorder (with the exception of major depressive or anxiety disorders), head injury, genetic disorder associated with autism (e.g., fragile X syndrome, tuberous sclerosis), or any other medical condition affecting brain function (e.g., epilepsy). All ASC participants were diagnosed with ASC according to ICD-10 research criteria. ASC diagnoses were confirmed using the Autism Diagnostic Interview-Revised (ADI-R) [40] and it was allowed for participants to be 1 point below cutoff for one of the three ADI-R domains in the diagnostic algorithm. The Autism Diagnostic Observation Schedule (ADOS) [41], was used to assess current symptoms for all participants with ASC. The Wechsler Abbreviated Scale of Intelligence (WASI) [42] was used to assess Verbal IQ (VIQ), Performance IQ (PIQ) and Full Scale IQ (FSIQ). Depression and

anxiety symptoms were also measured with the Beck Depression Inventory (BDI) and Beck Anxiety Inventory (BAI).

Reading the Mind in the Eyes Test (RMET)

All participants in both discovery and replication datasets completed the 'Reading the Mind in the Eyes' Test (RMET). The RMET [43] consists of 36 items of grey-scale photos cropped and rescaled so that only the area around the eyes can be seen. Each photo is surrounded by four mental state terms and the participant is instructed to choose the word that best describes what the person in the photo is thinking or feeling. Participants in both discovery and replication datasets completed a computerized online version of the RMET at home. Participants were instructed to select the most appropriate item within 20 seconds for each stimulus (presented in random order). Responses were coded as correct or incorrect (wrong items selected, or no response after 20 seconds), giving a maximum total correct score of 36. To guard against the possibility that many items timed-out, we used a rule that if an individual had time-outs on 9 or more items (>25% of all items), then such individuals were excluded from analysis. All participants in both discovery and replication datasets also completed the Autism Spectrum Quotient (AQ) [44] and the Empathy Quotient (EQ) [45] on the same online platform and before taking the RMET.

Statistical Analysis

Subgrouping was achieved using agglomerative hierarchical clustering on a data matrix of binary values (1 for correct, 0 for incorrect) with the dimensions of rows specifying subjects and columns specifying RMET items. These analyses were performed using the clustergram.m function within the Bioinformatics Toolbox of MATLAB R2015a, which performs clustering both on the rows (i.e. subjects) and columns (i.e. RMET items) of the data matrix. Ward's metric was used as the linkage method across both subjects and RMET items. The distance metric applied to both rows and columns was the Hamming distance and is operationalized as the percentage of items between individuals that differ and is appropriate for the context of binary data. To determine the number of subgroups present along the subject dimension, we used two separate cluster validity indices to evaluate the optimal number of clusters; pseudot2 and silhouette. The pseudot2 indice is only applicable for hierarchical clustering methods [46] and is implemented within the NbClust library in R [47]. For our purposes, we translated the R code from NbClust into MATLAB code in order to accommodate utilization of the Hamming distance metric we have applied in our analyses. The silhouette indice [48] is a measure of how similar a particular subject is with its own cluster compared subjects from other clusters. Silhouette values ranges from -1 to 1 and high silhouette values indicate that subjects within a cluster are well matched (i.e. highly similar) to that cluster and are poorly matched to other clusters. The optimal number of clusters is the cluster solution with the highest silhouette value. This method is

implemented within the MATLAB function evalclusters.m. To test for the consistency of such optimal cluster solutions, we used bootstrapping (1000 resamples) to examine the frequency with which the observed optimal cluster solution occurred.

To test the idea that unsupervised stratification of individuals in one dataset will lead to robust replicable detection of individuals within the same subgroups in a second independent dataset, we used a linear support vector machine (SVM) classifier implemented within LIBSVM (https://www.csie.ntu.edu.tw/~cjlin/libsvm/). The ASC Impaired subgroup was set to class 1 and the combination of ASC Unimpaired and TD individuals was set to class 2. The regularization parameter, C, was set to the default of 1. Because of the imbalance in class size (more ASC Unimpaired and TD individuals than ASC Impaired), we used a class-weighting scheme whereby class 1 was weighted as 1/n1 and class 2 was weighted as 1/n2, where n1 and n2 are the sizes of each class. We used two-fold cross-validation whereby on fold 1, the discovery set (CARD) was the training set and the replication set (AIMS) the test set. Fold 2 reversed this order (i.e., AIMS as training, CARD as test) and performance metrics of accuracy, sensitivity and specificity were averaged across the folds to get a final estimate of performance. Permutation tests (10,000 permutations) were implemented to compare observed performance levels against performance levels observed under the null distribution and to compute p-values for each performance metric.

Once ASC subgroups were defined, we computed total RMET scores (i.e. sum across all items) for each individual and ran independent samples t-tests to specifically compare the total score in ASC subgroups to the TD group. These t-statistics were also used to compute replication Bayes Factor (BF) statistics to quantify strength of evidence for replication (BF ~1 indicates little to no evidence supporting replication, BF>10 indicates strong evidence for replication, BF>100 indicates extremely strong evidence for replication). These replication BF statistics are computed with code accompanying a recent paper by Verhagen and Wagenmakers [49] (see here for R code: http://bit.ly/1GHiPRe).

We also examined other variables such as sex/gender, VIQ, age, AQ, EQ, BDI, BAI, and ADOS and ADI-R subscales to test hypotheses about whether the subgroups would differ on these variables. To test for the possibility of imbalances across the subgroups as a function of sex/gender, we counted up the number of males and females across all subgroups and compared them to expected counts derived from a chi-square test. To test VIQ, age, AQ, EQ, BDI, BAI, ADOS, and ADI-R score differences we used independent samples t-tests to compare measures between ASC subgroups. Given that VIQ emerged as significantly different across the subgroups, we re-ran statistical tests evaluating RMET between-group differences compared to TD with VIQ as a covariate. This test was implemented using the glmfit.m function in MATLAB.

Results

Clustering of the ASC group within the discovery dataset (CARD) is shown in Fig 1A. Two distinct ASC subgroups emerged as the two main branches of the dendrogram and this two-cluster solution was confirmed by both the pseudot2 and silhouette cluster validity indices. Similarly in the replication dataset (AIMS), there was also evidence of two distinct ASC subgroups again confirmed as the optimal number of clusters by both pseudot2 and silhouette cluster validity indices (Fig 1B). Bootstrapping to examine the consistency of these two-cluster solutions showed that they were the predominant solution across 1000 bootstrap resamples (percentage of bootstrap resamples with a two-cluster solution: Discovery pseudot2 = 100%; Discovery silhouette = 100%; Replication pseudot2 = 100%; Replication silhouette = 59%). These results indicate that adults with ASC can be stratified into at least 2 subgroups, which can be generally characterized as 'Impaired' and 'Unimpaired' mentalizing subgroups. The 'Impaired' subgroup comprised about 15% of ASC cases in the discovery set and 33% of ASC cases in the replication set, and could be descriptively characterized as showing a large proportion of incorrect items across all subjects (indicated in Fig 1 as black cells within the clustergrams representing data across all subjects and items). The 'Unimpaired' subgroup comprised 85% of ASC cases in the discovery set and 67% of ASC cases in the replication set, and could be descriptively characterized as showing a large number of correct items across all subjects (indicated in Fig 1 as red cells within the clustergrams representing data across all subjects and items).

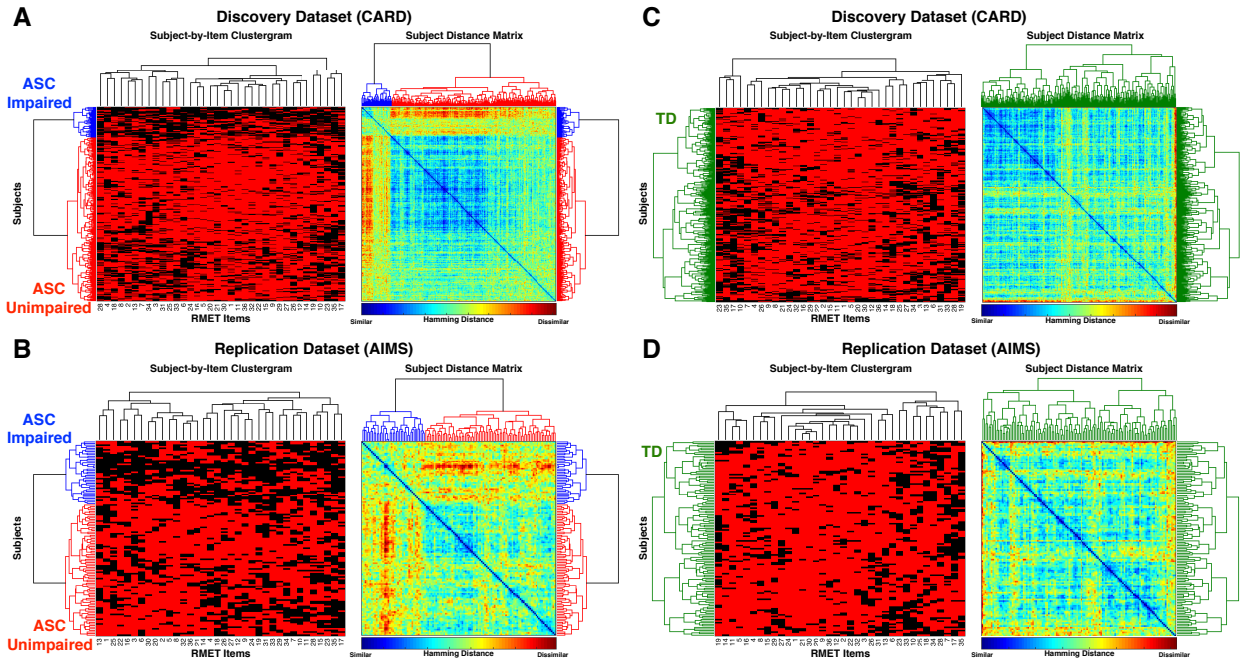


Fig. 1: Clustering individuals and RMET items across discovery and replication datasets. *This figure shows clustergrams (i.e. two-way hierarchical clustering across subjects (rows) and*

RMET items (columns)) and distance matrices for both ASC (A, C) and TD (B, D) groups in the discovery (A, C) and replication (B, D) datasets (ASC Impaired, blue; ASC Unimpaired, red; TD, green). Within the clustergrams, the cells colored in black denote RMET items where the response was incorrect, whereas cells colored in red denote items where the response was correct. To the right of the subject-by-item clustergrams are distance matrices depicted as a heat map. These distance matrices show the Hamming distance between all subjects. The rows and columns of the distance matrices are ordered in the same way as the dendrogram for hierarchical clustering. The color values indicate similarity between subjects with cool colors indicating high similarity whereas hot colors indicate high dissimilarity.

Using classification analyses we were able to test the idea that such subgroups are replicable, since training on such subgroup distinctions in one dataset should lead to generalizable accurate predictions of subgroup membership in an independent dataset. A linear SVM classifier performs with 92.16% accuracy ($p = 0.0012$), 89.05% sensitivity ($p = 0.0077$), and 92.89% specificity ($p = 2.99e\text{-}4$) for distinguishing the ASC Impaired subgroup from all other individuals (ASC Unimpaired and TD). Similarly, if one ignores the TD data completely and only focuses on the distinction between ASD Impaired versus ASD Unimpaired, performance was also high (Accuracy = 89.54%, $p = 3.99e\text{-}4$; Sensitivity = 88.66%, $p = 4.99e\text{-}4$; Specificity = 90.49%, $p = 1.99e\text{-}4$).

Underscoring these analyses, we have also computed dissimilarity between-subjects on a within- or between-dataset basis. These results can be seen in Fig 2. In Fig 2C, we show the critical comparisons of how similarity in RMET item-level responses are higher within ASC subgroup but across datasets compared to between ASC subgroups within the same dataset. Here, we computed the average between-subject distance for each subject in relation to all other subjects in different datasets and subgroups. The idea here is that between-subject distance should be smaller within-subgroup but across dataset (i.e. subsets 5 and 6) than when comparing between-subject distance between-subgroups but within the same dataset (i.e. subsets 7 and 8). The effect size difference for all of these comparisons (i.e. subset 5 vs 7 or 8 and subset 6 vs 7 or 8) was always greater than 0.8, indicating indeed that at a quantitative level there is much similarity in RMET item-level responses within individuals of the same subgroup, even when looking across different datasets.

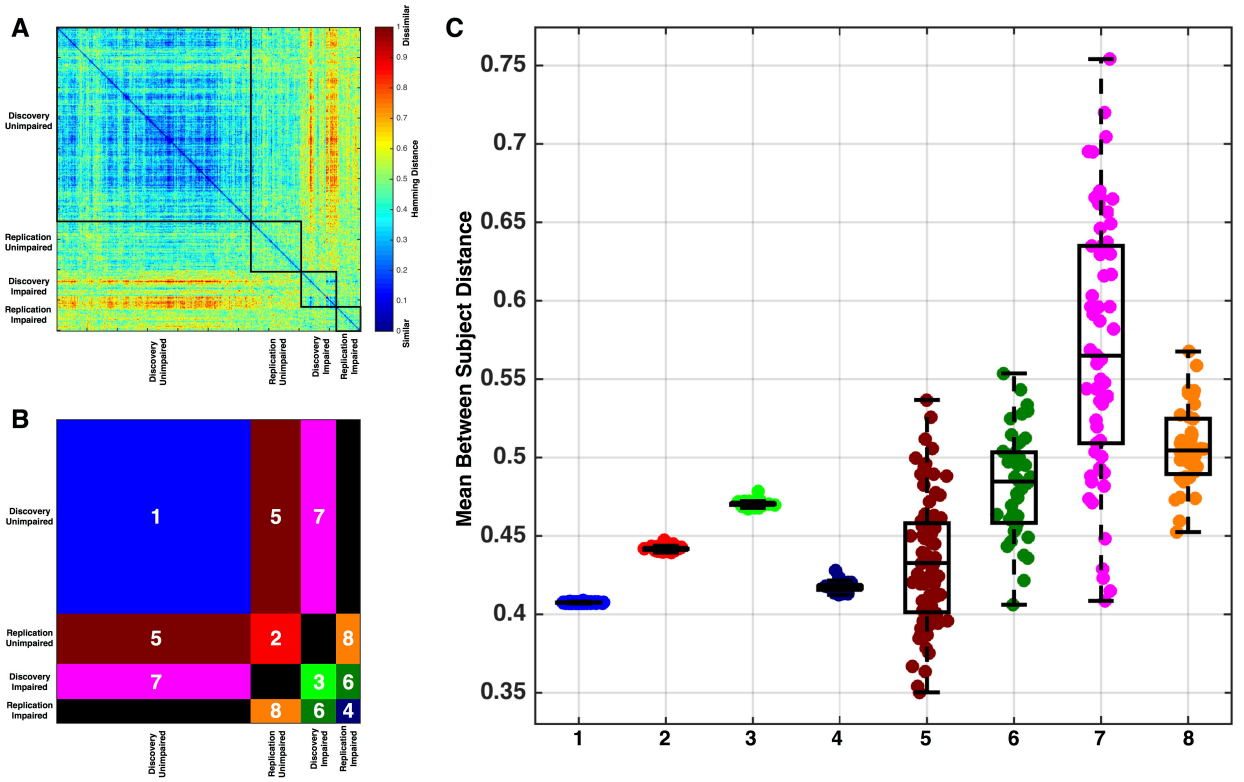


Fig. 2: Quantification of between-subject dissimilarity in RMET item-level responses within ASC subgroups and between datasets. *Panel A shows a dissimilarity matrix of Hamming Distance for all ASC subjects across Discovery and Replication datasets. The subgroups within each dataset are demarcated with the black outlines. Cells outside of the black outlines depict between-subject dissimilarity (i.e. Hamming distance) across subgroups and datasets. Panel B shows color and number markings for each subset of data shown in Panel C. Panel C shows the mean between-subject dissimilarity computed on an individual subject basis. This means each dot in panel C represents how that one individual on-average differs from all other individuals within a particular subset. The critical comparisons showing that RMET item-level responses are more similar within ASC subgroup across datasets than across ASC subgroups but within the same dataset are subsets 5 and 6 compared to subsets 7 and 8. Comparing subsets 5 or 6 to 7 or 8 yields an effect size greater than Cohen's d = 0.8.*

Clustering on the TD group across both datasets also resulted in two-cluster solutions across both pseudot2 and silhouette cluster validity indices, and these two-cluster solutions were highly consistent across bootstrap resamples (96-100%) (see Supplementary Figure 1). However, the clusters across each dataset appeared different. Within the discovery dataset, the two TD subgroups appeared to be different in terms of overall correct/incorrect scores, with one of the subgroups scoring lower than the other ($t(314)$ = -12.24, p = 1.86e-28, *Cohen's d* = 1.42). In contrast, within the replication dataset, the two TD subgroups were not different in total scores ($t(124)$ = -1.50, p = 0.13, *Cohen's d* = 0.27), which indicates that the primary difference between the subgroups is one of different patterning of response across RMET items, but with no overall

effect on total RMET score. Confirming the idea that the TD subgroups across datasets are different, an SVM classifier trained on one dataset to predict individuals in subgroups from the other dataset performed poorly (Accuracy = 53.95%, Sensitivity = 58.21%, Specificity = 49.98%). Because there was no clear replicable distinction across datasets for these TD subgroups, we defaulted on a one-cluster solution (i.e., no clear subgroup distinction) in all further analyses that compare TD to ASC subgroups (Fig 1C-D). However, this does not imply that subgroups don't exist within the TD population. Rather, it may simply be that our sample sizes in each dataset are too small to pick up on subtle replicable subgroup differentiation.

Given the robust presence of ASC subgroups, we next compared these subgroups to TD using RMET summary scores. These tests allow us to examine how the subgroups compare to TD in mentalizing ability. Here we find that the 'Unimpaired' subgroup was not differentiated from the TD group in the discovery set (Discovery: $t(634) = 0.13$, $p = 0.89$, *Cohen's d* = 0.01) and this effect replicates in the replication set (Replication: $t(207) = -5.51$, $p = 4.22e\text{-}7$, *Cohen's d* = 0.73; *BF* = 16.16). Note that the replication Bayes Factor (BF) was greater than 10, which generally indicates strong evidence in favor of replication. Since there was no effect in the discovery set, this BF can be interpreted as replication of no true difference. However, one caveat to keep in mind is that the dataset showing little to no impairment (i.e. discovery) is a dataset where ASC diagnosis is self-reported and verification of such diagnosis was not possible. In contrast, the dataset where ASC diagnosis is validated with gold-standard instruments, we do see some evidence of reduced scores in this subgroup. Therefore, a cautious and perhaps conservative interpretation of this subgroup would be to describe it as an 'Unimpaired' subgroup because of the lack of replicable evidence of difficulty in mentalizing, but to also state that a better test of replication would likely be on two datasets with similar and strict criteria for inclusion of ASC individuals.

In contrast to the 'Unimpaired' subgroup, the 'Impaired' subgroup showed massively reduced performance on the RMET compared to the TD group, with the effect size being greater than 3 standard deviations in both datasets (Discovery: $t(372) = -21.39$, $p = 8.40e\text{-}67$, Cohen's d = 3.05; Replication: $t(164) = -18.47$, $p = 6.27e\text{-}42$, *Cohen's d* = 3.35). Furthermore, this effect was highly replicable across datasets (*BF* = 7.40e+39), albeit again with a similar pattern of the effect size of the replication set being larger than that of the discovery set. We further tested whether this effect of reduced RMET scores in the Impaired group would remain in the Replication set after covarying for sex/gender, age, center, and VIQ. After controlling for these factors, RMET scores were still reduced in the ASC Impaired subgroup versus the TD group ($t(159) = -17.25$; $p = 2.91e\text{-}38$) and retained evidence for replicability (*BF* = 6.73e+36). These results indicate that our subgrouping strategy effectively enhances sensitivity for highlighting a subgroup with impaired mentalizing ability in adulthood and also increases specificity by showing that only one ASC subgroup is impaired while the other subgroup shows intact mentalizing ability.

Within each dataset, it would not be statistically appropriate to compare the Impaired vs Unimpaired ASC subgroups directly, since the selection process of clustering subgroups and the subsequent hypothesis test are both done on the same data [50, 51]. However, we can compare the Impaired subgroup from one dataset to the Unimpaired subgroup from a separate independent dataset, because here the selection process is independent of the hypothesis test. Comparing the discovery set 'Impaired' subgroup to the replication set 'Unimpaired' subgroup, we found that the Impaired subgroup showed massively reduced performance on the RMET ($t(139)$ = -18.05, p = 2.83e-38, *Cohen's d* = 3.09). Conversely, comparison of the discovery 'Unimpaired' subgroup with replication 'Impaired' subgroup also indicated massively reduced scores in the 'Impaired' subgroup ($t(358)$ = 16.04, p = 5.02e-44, *Cohen's d* = 2.69). Computation of a replication Bayes Factor indicated that this effect was highly replicable (BF = 1.24e+36).

To further facilitate visualization of these subgroups and how such a subgrouping approach parsimoniously breaks down the variability within an omnibus ASC group, we have plotted summary scores for the ASC subgroups separately in Fig 3A-B and again in Fig 3C-D when the ASC individuals are combined as one omnibus group. It is clear from these plots that these subgroups cannot be simply derived by using a hard threshold on RMET summary scores, and this illustrates how the subgrouping approach via hierarchical clustering of item-level responses on the RMET can go beyond limitations apparent when one utilizes summary scores alone. Also of note is the enhanced sensitivity of documenting mentalizing difficulties in ASC, when ASC Impaired is compared to TD versus when the omnibus ASC group is compared to TD.

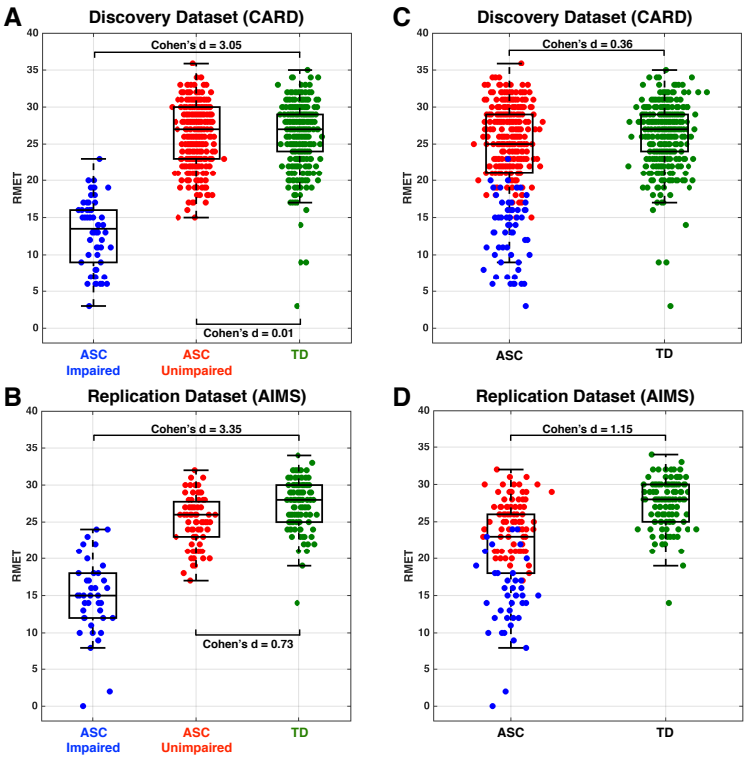


Fig. 3: RMET total scores across subgroups in discovery and replication datasets. *RMET total scores are shown as boxplots along with dots representing individual data points in discovery (A, C) and replication (B, D) datasets (ASC Impaired, blue; ASC Unimpaired, red; TD, green). Panels A and B show the data when subgroups are separated while panels C and D show the data when ASC is treated as one omnibus group.*

Examination of other variables such as sex/gender, age, autistic traits, trait empathy, depression and anxiety symptoms, and autism symptom severity showed no signs of consistent differentiation across ASC subgroups (see the supplementary text and figures). However, it is noteworthy that within the Discovery dataset, there was evidence for higher autistic traits on the AQ and lower trait empathy on the EQ (Supplementary Figures 5A and 6A). Aside from these other measures, the only other measure showing signs of differentiation amongst the subgroups was VIQ. VIQ was reduced in the ASC Impaired subgroup compared to the ASC Unimpaired and TD groups. This effect of reduced VIQ in the ASC Impaired subgroup did not change the main inferences with regards to ASC Impaired showing reduced RMET total scores, as shown by the earlier analysis treating VIQ as a covariate.

Discussion

In this study we examined the issue of heterogeneity in mentalizing ability in adults with ASC. We discovered distinct, replicable, and robust data-driven ASC subgroups who differ in performance on the RMET. The first subgroup, whom we call the 'Impaired' subgroup, consists of only a small subset of ASC adults (15 to 33%) and exhibits profound difficulties within the domain of mentalizing with heavily reduced RMET scores compared to typically-developing controls. In contrast, the second subgroup, whom we call the 'Unimpaired' subgroup, consists of a majority of ASC adults (67 to 85%) and shows little to no consistent deficit in mentalizing as measured by the RMET. Parsing ASC mentalizing heterogeneity into these subgroups is important for several reasons. First, rather than a majority of individuals showing difficulty, only a minority of adults with ASC still show substantial difficulty in mentalizing in adulthood. In the context of standard case-control studies that look for on-average group differences, it is clear that this small subset is primarily driving the effects observed in adults on the RMET [37]. Given this insight, it seems like an obvious necessity for future work to take a similar stratification approach as we have, in order to break through with more clear and precise interpretations about who possesses mentalizing deficits in adults with ASC via the RMET.

Of equal importance is the idea that a majority of adults with ASC show little to no difficulty on this advanced mentalizing task. This could potentially mean that mentalizing ability for a majority of ASC adults has developed into an effective ability that can be explicitly deployed in specific circumstances. However, although we are describing this subgroup comprising a majority of ASC adults as having little to no difficulty, an important caveat is that when ASC individuals are strictly included after gold-standard instrument diagnostic validation (i.e. replication dataset) as opposed to simply self-reporting ASC diagnosis (i.e. discovery dataset), there was a notable decrease in RMET scores in this group (i.e. Cohen's d = 0.73). This could mean that a dataset with unverified self-diagnoses could underestimate the true subtle difficulty this subgroup might have in mentalizing ability. In other words the effect size estimate which may be more indicative of the true effect size might be that of the replication set simply

because diagnoses were all verified with gold-standard instruments. It is also possible that this result reflects a subtle difficulty in aspects of mentalizing in this subgroup, but that the RMET may not be as sensitive to picking up such difficulty. It may be that more sensitive tests looking at understanding in real-world naturalistic social interactions, or which tap an individual's ability to automatically or implicitly mentalize, or which employ utilization of mentalizing to make moral judgments, could all pull apart subtle yet persisting deficits in mentalizing [6, 19, 27, 28, 52-55]. It will be useful for future work to specifically examine this 'Unimpaired' group to test whether more sensitive tests could reveal more subtle (or other aspects of) deficits or whether this group could indeed be characterized as completely unimpaired in mentalizing ability. Nevertheless, our findings of discrete, replicable, and robust ASC subgroups with differing mentalizing ability in adulthood represents an important stride forward in terms of the precision of our understanding of mentalizing difficulties in adulthood in individuals with ASC. This work is highly compatible with the goals of 'precision medicine' or 'stratified psychiatry' and is what is needed to move forward with research that has clinical impact for patients and which can also further translational research progress focused on honing in on treatment-relevant mechanisms [29, 30].

In addition to the clinical impact of such insights about mentalizing heterogeneity, the current study could have potentially large impact on basic areas of research on ASC. For example, inconsistency within the functional and structural neuroimaging literature on ASC (e.g., [16, 23, 56, 57]) could be mitigated by a better understanding of mentalizing heterogeneity nested within relatively small ASC samples typically utilized in such work. This point is exemplified by recent work by Byrge and colleagues, whereby it is suggested that some group-level differences in case-control designs could be driven by the effects nested within a small subgroup of patients [28]. A better a priori understanding of the heterogeneity present within the ASC population could be of large impact for study design and could also implicate different underlying mechanisms that explain such heterogeneity (e.g., [31, 58-60]).

These subgroup distinctions are also particularly important to make because of how they apply specifically to the RMET. The RMET is a long-standing standardized instrument that is widely used within the autism literature and within social neuroscience in general. The NIMH Research Domain Criteria (RDoC) lists the RMET as one of several important tests for characterizing variation in social processes, particularly under the category of Perception and Understanding of Others (http://1.usa.gov/1Qs6MdI). With regards to treatment research, the RMET is widely used as treatment outcome measure, particularly for drug manipulations (i.e. oxytocin) or behavioral interventions targeting social skills and social cognition (e.g., [32-35]). All of this prior work utilizes an analytic strategy of computing RMET summary scores across all items and then onto potentially sub-optimal omnibus case-control comparisons which mask the presence of nested subgroups within ASC. The current work should signal a change in this practice for how the RMET is utilized in important clinical settings (e.g., evaluating treatment outcome). Rather than using summary scores in an omnibus ASC group, a more fruitful approach

would be to use the RMET to distinguish 'Impaired' vs 'Unimpaired' subgroups and to then specifically evaluate whether such ASC subgroups respond differently to treatment. In other words, the added knowledge we provide here is that these subgroups could signal a meaningful distinction that helps the design of intervention studies and the interpretation of treatment findings. Given the current state of largely mixed results for most interventions for ASC [36], it may become clearer after subgrouping that some existing treatments do systematically work for particular subgroups but not others.

The approach we have taken to subgroup the autism spectrum is also worth highlighting in terms of its novelty and advantages. Rather than utilizing the RMET in a standard approach by summarizing all items into one total score, we have instead retained the full data set of information encoded across the 36 items as input into *unsupervised* hierarchical clustering algorithms that came to data-driven conclusions about the presence of 2 distinct ASC subgroups. This unsupervised approach avoids using experimenter-derived cutoffs and instead is completely utilizing data-driven distinctions that are robust enough to emerge in a replicable fashion across independent datasets. Data-driven subgrouping is by no means specific to the RMET. For example, recent work has applied similar logic to clustering the phenotype based on gold-standard diagnostic instruments [61] and for clustering 26 different mouse models of genetic mechanisms related to ASC [58]. Clustering also forms the bedrock of systems biology genomic approaches such as gene co-expression network analysis [62, 63], which is a highly utilized approach in autism genetics research [64-68]. Subgrouping approaches like clustering are generalizable to any measure, and could be used in a whole range of new applications focused on data-driven stratification in ASC.

A further advantage to our approach of finding data-driven distinctions is that such distinctions are generalizable across datasets. As we have shown with the classification analyses, the stratifications made in one dataset generalize to highly accurate predictions of the same data-driven subgroups in independent data. In future work, such information about replicable subgroups could be turned into valuable assessment or research tools that could aid in study design and participant/patient screening. For instance, randomized control trials may screen patients along such distinctions or may use the RMET as an outcome measure and use such distinctions to analyze individualized treatment response patterns. Such stratifications could also be useful in clinical assessments and facilitates individualized treatment planning.

In addition to highlighting the promise of such stratifications for autism research, there are also caveats that we have assessed in further follow-up analyses. We have found that ASC subgroups are not systematically differentiated as a function of sex/gender, age, autistic traits, trait empathy or symptom severity measured by ADI-R and ADOS. While most of these measures showed consistency across Discovery and Replication datasets in the lack of subgroup differentiation, the exceptions were the AQ and EQ. For both AQ and EQ, we found that

although the Discovery dataset showed a difference of higher autistic traits and lower empathy in the ASC Impaired versus Unimpaired subgroup, this effect was not observed in the Replication dataset. This lack of consistency may indicate differences in sampling and diagnostic verification across the datasets may be relevant. In addition, the much larger Discovery dataset (but not the Replication dataset) shows notably long-tailed distributions for the Unimpaired subgroup on both measures (Supplementary Figures 5A and 6A). This characteristic of the ASC Unimpaired distributions may be one that may become more apparent in cohorts of much larger sizes and could potentially explain why such differences appear for the Discovery but not Replication dataset. At the cognitive level of explanation, it could be that the Unimpaired subgroup is potentially still a heterogeneous mixture of individuals that cannot be easily parsed via the RMET, as some individuals in this group may be truly unimpaired, while others may mask deficits via other compensatory mechanisms that allow them to 'hack' out explicit solutions in ways that TD individuals may not utilize. Future work employing more sensitive mentalizing measures as well as other sensitive measures of everyday social functioning may be able to pick apart these aspects of heterogeneity in the Unimpaired subgroup. And finally, a third alternative explanation could be related to individual differences in self-referential cognitive ability such as self-insight that both self-report AQ and EQ measures rest upon. Theoretically, there are very important links between self-referential cognition and the domain of mentalizing [7-14], and this may ultimately influence the accuracy of self-report for some ASC patients [69], but may also be integrally related to the difficulties some patients have in the domain of mentalizing.

Perhaps hinting at some level of non-social cognitive ability that differentiates these RMET-defined ASC subgroups, we discovered that the Impaired subgroup is lower in VIQ than both ASC Unimpaired and TD groups. This VIQ effect on the RMET has been noted before [70] and may be easily understood given that the RMET may tax vocabulary for some individuals and on some items. Despite this effect of VIQ, we found that the reduction in mentalizing ability in the ASC Impaired subgroup could not be accounted for simply by this VIQ effect. When controlling for the effect of VIQ as well as a range of other factors (e.g., sex/gender, age, center) we still found evidence for robust deficits in mentalizing within the ASC Impaired subgroup. This evidence suggests that while some variability in RMET performance is linked to variability in VIQ, these subgroup distinctions are not fully explained by VIQ variation. Rather, the ASC Impaired subgroup consists of individuals with particular difficulty in mentalizing in adulthood.

In conclusion, the discoveries in this study allow for a more precise understanding of mentalizing difficulties in adults with ASC. Our insights have the potential to further personalized medicine aims in ways that accelerate progress towards clinical impact for patients. By understanding how the autism spectrum can be stratified in clinically meaningful ways, translational opportunities also may open up that could test whether such distinctions are rooted in distinct underlying mechanisms.

Acknowledgments

This study was supported by the UK Medical Research Council (MRC), Wellcome Trust and the Autism Research Trust. This study was conducted in association with the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care East of England at Cambridgeshire and Peterborough NHS Foundation Trust (NIHR CLHARC EoE) and the European Autism Interventions—A Multicentre Study for Developing New Medications (EU-AIMS) consortium; EU-AIMS receives support from the Innovative Medicines Initiative Joint Undertaking under grant agreement number 115300, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013), EFPIA companies, and Autism Speaks. M-CL and AR were supported by the William Binks Autism Neuroscience Fellowship at the University of Cambridge. M-CL is supported by the O'Brien Scholars Program within the Child and Youth Mental Health Collaborative at the Centre for Addiction and Mental Health and The Hospital for Sick Children, Toronto.

References

1. Baron-Cohen S, Leslie AM, Frith U. Does the autistic child have a "theory of mind"? *Cognition* 1985; **21**(1)**:** 37-46.

2. Baron-Cohen S. *Mindblindness: An essay on autism and theory of mind.* MIT Press: Cambridge, MA, 1995.

3. Frith U. Mind blindness and the brain in autism. *Neuron* 2001; **32**(6)**:** 969-979.

4. Frith U, Happe F, Siddons F. Autism and theory of mind in everyday life. *Social Development* 1994; **3**(2)**:** 108-124.

5. Happe FG. The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Dev* 1995; **66**(3)**:** 843-855.

6. Senju A, Southgate V, White S, Frith U. Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome. *Science (New York, NY* 2009; **325**(5942)**:** 883-885.

7. Lombardo MV, Barnes JL, Wheelwright SJ, Baron-Cohen S. Self-referential cognition and empathy in autism. *PLoS One* 2007; **2**(9)**:** e883.

8. Lombardo MV, Baron-Cohen S. Unraveling the paradox of the autistic self. *WIREs Cognitive Science* 2010; **1:** 393-403.

9. Lombardo MV, Baron-Cohen S. The role of the self in mindblindness in autism. *Consciousness and cognition* 2011; **20**(1)**:** 130-140.

10. Williams D. Theory of own mind in autism: Evidence of a specific deficit in self-awareness? *Autism : the international journal of research and practice* 2010; **14**(5)**:** 474-494.

11. Williams DM, Happe F. What did I say? Versus what did I think? Attributing false beliefs to self amongst children with and without autism. *J Autism Dev Disord* 2009; **39**(6)**:** 865-873.

12. Frith U. *Autism: Explaining the enigma.*, 2nd edn. Blackwell: Malden, MA, 2003.

13. Frith U, Happe F. Theory of mind and self-consciousness: What it is like to be autistic. *Mind and Language* 1999; **14:** 1-22.

14. Hurlburt RT, Happe F, Frith U. Sampling the form of inner experience in three adults with Asperger syndrome. *Psychological medicine* 1994; **24**(2)**:** 385-395.

15. Castelli F, Frith C, Happe F, Frith U. Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain* 2002; **125**(Pt 8)**:** 1839-1849.

16. Lombardo MV, Chakrabarti B, Bullmore ET, Consortium MA, Baron-Cohen S. Specialization of right temporo-parietal junction for mentalizing and its association with social impairments in autism. *Neuroimage* 2011; **56:** 1832-1838.

17. White SJ, Frith U, Rellecke J, Al-Noor Z, Gilbert SJ. Autistic adolescents show atypical activation of the brain's mentalizing system even without a prior history of mentalizing problems. *Neuropsychologia* 2014; **56:** 17-25.

18. Baron-Cohen S, Ring HA, Wheelwright S, Bullmore ET, Brammer MJ, Simmons A*, et al*. Social intelligence in the normal and autistic brain: An fMRI study. *The European journal of neuroscience* 1999; **11:** 1891-1898.

19. Pantelis PC, Byrge L, Tyszka JM, Adolphs R, Kennedy DP. A specific hypoactivation of right temporo-parietal junction/posterior superior temporal sulcus in response to socially awkward situations in autism. *Soc Cogn Affect Neurosci* 2015; **10**(10)**:** 1348-1356.

20. Redcay E, Dodell-Feder D, Mavros PL, Kleiner M, Pearrow MJ, Triantafyllou C*, et al*. Atypical brain activation patterns during a face-to-face joint attention game in adults with autism spectrum disorder. *Hum Brain Mapp* 2013; **34**(10)**:** 2511-2523.

21. Kana RK, Keller TA, Cherkassky VL, Minshew NJ, Just MA. Atypical frontal-posterior synchronization of Theory of Mind regions in autism during mental state attribution. *Soc Neurosci* 2009; **4**(2)**:** 135-152.

22. Kana RK, Libero LE, Hu CP, Deshpande HD, Colburn JS. Functional brain networks and white matter underlying theory-of-mind in autism. *Soc Cogn Affect Neurosci* 2014; **9**(1)**:** 98-105.

23. Dufour N, Redcay E, Young L, Mavros PL, Moran JM, Triantafyllou C*, et al*. Similar brain activation during false belief tasks in a large sample of adults with and without autism. *PLoS One* 2013; **8**(9)**:** e75468.

24. Lai MC, Lombardo MV, Baron-Cohen S. Autism. *Lancet* 2014; **383:** 896-910.

25. Lai MC, Lombardo MV, Chakrabarti B, Baron-Cohen S. Subgrouping the autism "spectrum": reflections on DSM-5. *PLoS biology* 2013; **11**(4)**:** e1001544.

26. Brock J. Commentary: Complementary approaches to the developmental cognitive neuroscience of autism--reflections on Pelphrey et al. (2011). *J Child Psychol Psychiatry* 2011; **52**(6)**:** 645-646.

27. Moran JM, Young LL, Saxe R, Lee SM, O'Young D, Mavros PL*, et al*. Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences of the United States of America* 2011; **108**(7)**:** 2688-2692.

28. Byrge L, Dubois J, Tyszka JM, Adolphs R, Kennedy DP. Idiosyncratic brain activation patterns are associated with poor social comprehension in autism. *J Neurosci* 2015; **35**(14)**:** 5837-5850.

29. Kapur S, Phillips AG, Insel TR. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular psychiatry* 2012; **17**(12)**:** 1174-1179.

30. Insel TR. The NIMH Research Domain Criteria (RDoC) Project: precision medicine for psychiatry. *The American journal of psychiatry* 2014; **171**(4)**:** 395-397.

31. Lombardo MV, Pierce K, Eyler LT, Carter Barnes C, Ahrens-Barbeau C, Solso S*, et al*. Different functional neural substrates for good and poor language outcome in autism. *Neuron* 2015; **86**(2)**:** 567-577.

32. Soorya LV, Siper PM, Beck T, Soffes S, Halpern D, Gorenstein M*, et al*. Randomized comparative trial of a social cognitive skills group for children with autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry* 2015; **54**(3)**:** 208-216 e201.

33. Anagnostou E, Soorya L, Chaplin W, Bartz J, Halpern D, Wasserman S*, et al*. Intranasal oxytocin versus placebo in the treatment of adults with autism spectrum disorders: a randomized controlled trial. *Mol Autism* 2012; **3**(1)**:** 16.

34. Einfeld SL, Smith E, McGregor IS, Steinbeck K, Taffe J, Rice LJ*, et al*. A double-blind randomized controlled trial of oxytocin nasal spray in Prader Willi syndrome. *Am J Med Genet A* 2014; **164A**(9)**:** 2232-2239.

35. Cacciotti-Saija C, Langdon R, Ward PB, Hickie IB, Scott EM, Naismith SL*, et al*. A double-blind randomized controlled trial of oxytocin nasal spray and social cognition training for young people with early psychosis. *Schizophr Bull* 2015; **41**(2)**:** 483-493.

36. Fletcher-Watson S, McConnell F, Manola E, McConachie H. Interventions based on the Theory of Mind cognitive model for autism spectrum disorder (ASD). *The Cochrane database of systematic reviews* 2014; **3:** CD008785.

37. Baron-Cohen S, Bowen DC, Holt RJ, Allison C, Auyeung B, Lombardo MV*, et al*. The "Reading the Mind in the Eyes" Test: Complete Absence of Typical Sex Difference in ~400 Men and Women with Autism. *PLoS One* 2015; **10**(8)**:** e0136521.

38. Daniels AM, Rosenberg RE, Anderson C, Law JK, Marvin AR, Law PA. Verification of parent-report of child autism spectrum disorder diagnosis to a web-based autism registry. *J Autism Dev Disord* 2012; **42**(2)**:** 257-265.

39. Wilson CE, Happe F, Wheelwright SJ, Ecker C, Lombardo MV, Johnston P*, et al*. The neuropsychology of male adults with high-functioning autism or asperger syndrome. *Autism Res* 2014; **7**(5)**:** 568-581.

40. Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord* 1994; **24**(5)**:** 659-685.

41. Lord C, Risi S, Lambrecht L, Cook EH, Jr., Leventhal BL, DiLavore PC*, et al*. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J Autism Dev Disord* 2000; **30**(3)**:** 205-223.

42. Wechsler D. *Wechsler abbreviated scale of intelligence.* The Psychological Corporation: San Antonio, TX, 1999.

43. Baron-Cohen S, Wheelwright S, Hill J, Raste Y, Plumb I. The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J Child Psychol Psychiatry* 2001; **42**(2)**:** 241-251.

44. Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E. The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *J Autism Dev Disord* 2001; **31**(1)**:** 5-17.

45. Baron-Cohen S, Wheelwright S. The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *J Autism Dev Disord* 2004; **34**(2)**:** 163-175.

46. Duda RO, Hart PE. *Pattern classification and scene analysis*, vol. 3. Wiley: New York, 1973.

47. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software* 2014; **61:** 1-36.

48. Rousseeuw P. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics* 1987; **20:** 53–65.

49. Verhagen J, Wagenmakers EJ. Bayesian tests to quantify the result of a replication attempt. *J Exp Psychol Gen* 2014; **143**(4)**:** 1457-1475.

50. Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience* 2009; **12**(5)**:** 535-540.

51. Vul E, Harris C, Winkielman P, Pashler H. Puzzlingly high correlations in fMRI studies of emotion, personality and social cognition. *Perspect Psychol Sci* 2009; **4:** 274-290.

52. Dziobek I, Fleck S, Kalbe E, Rogers K, Hassenstab J, Brand M, *et al*. Introducing MASC: a movie for the assessment of social cognition. *J Autism Dev Disord* 2006; **36**(5)**:** 623-636.

53. Golan O, Baron-Cohen S, Hill JJ, Golan Y. The "reading the mind in films" task: complex emotion recognition in adults with and without autism spectrum conditions. *Soc Neurosci* 2006; **1**(2)**:** 111-123.

54. Barnes JL, Lombardo MV, Wheelwright S, Baron-Cohen S. Moral dilemmas film task: A study of spontaneous narratives by individuals with autism spectrum conditions. *Autism research : official journal of the International Society for Autism Research* 2009; **2**(3)**:** 148-156.

55. Auyeung B, Lombardo MV, Heinrichs M, Chakrabarti B, Sule A, Deakin JB, *et al*. Oxytocin increases eye contact during a real-time, naturalistic social interaction in males with and without autism. *Transl Psychiatry* 2015; **5:** e507.

56. Haar S, Berman S, Behrmann M, Dinstein I. Anatomical Abnormalities in Autism? *Cereb Cortex* 2014.

57. Lefebvre A, Beggiato A, Bourgeron T, Toro R. Neuroanatomical Diversity of Corpus Callosum and Brain Volume in Autism: Meta-analysis, Analysis of the Autism Brain Imaging Data Exchange Project, and Simulation. *Biol Psychiatry* 2015; **78**(2)**:** 126-134.

58. Ellegood J, Anagnostou E, Babineau BA, Crawley JN, Lin L, Genestine M, *et al*. Clustering autism: using neuroanatomical differences in 26 mouse models to gain insight into the heterogeneity. *Molecular psychiatry* 2015; **20**(1)**:** 118-125.

59. Bernier R, Golzio C, Xiong B, Stessman HA, Coe BP, Penn O, *et al*. Disruptive CHD8 mutations define a subtype of autism early in development. *Cell* 2014; **158**(2)**:** 263-276.

60. Pickles A, Anderson DK, Lord C. Heterogeneity and plasticity in the development of language: a 17-year follow-up of children referred early for possible autism. *J Child Psychol Psychiatry* 2014; **55**(12)**:** 1354-1362.

61. Hu VW, Steinberg ME. Novel clustering of items from the Autism Diagnostic Interview-Revised to define phenotypes within autism spectrum disorders. *Autism Res* 2009; **2**(2)**:** 67-77.

62. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008; **9:** 559.

63. Parikshak NN, Gandal MJ, Geschwind DH. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat Rev Genet* 2015; **16**(8)**:** 441-458.

64. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S*, et al*. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 2011; **474**(7351)**:** 380-384.

65. Gupta S, Ellis SE, Ashar FN, Moes A, Bader JS, Zhan J*, et al*. Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat Commun* 2014; **5:** 5748.

66. Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V*, et al*. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 2013; **155**(5)**:** 1008-1021.

67. Pramparo T, Pierce K, Lombardo MV, Carter Barnes C, Marinero S, Ahrens-Barbeau C*, et al*. Prediction of autism by translation and immune/inflammation coexpressed genes in toddlers from pediatric community practices. *JAMA Psychiatry* 2015; **72**(4)**:** 386-394.

68. Pramparo T, Lombardo MV, Campbell K, Carter Barnes C, Marinero S, Solso S*, et al*. Cell cycle networks link gene expression dysregulation, mutation, and brain maldevelopment in autistic toddlers. *Mol Syst Biol* 2015; **11:** 841.

69. Griffin C, Lombardo MV, Auyeung B. Alexithymia in children with and without autism spectrum disorders. *Autism Res* 2015.

70. Peterson E, Miller SF. The Eyes Test as a Measure of Individual Differences: How much of the Variance Reflects Verbal IQ? *Front Psychol* 2012; **3:** 220.

Supplementary Text

Sex/Gender across ASC subgroups

The distribution of males and females across ASC subgroups was relatively equal, indicating no systematic bias for either males or females being over-/under-represented in any of the subgroups (Discovery: $\chi 2 = 1.27$, $p = 0.25$; Replication: $\chi 2 = 0.46$, $p = 0.49$; See Supplementary Figure 2).

Verbal IQ differences across ASC subgroups

Examination of VIQ (available only in the replication dataset) revealed evidence for slightly lower VIQ in the 'Impaired' subgroup compared to the 'Unimpaired' subgroup and the TD group, as well as lower VIQ in the Unimpaired versus TD group ($F(2,246) = 7.72$, $p = 5.59e\text{-}4$; 'Impaired' mean = 103.60, 'Unimpaired' mean = 111.36, TD mean = 113.13; 'Impaired' vs 'Unimpaired': $t(121) = -2.84$, $p = 0.005$, *Cohen's d* = 0.54; 'Impaired' vs TD: $t(164) = -5.71$, $p = 5.16e\text{-}8$, *Cohen's d* = 1.03; 'Unimpaired' vs TD: $t(207) = -4.95$, $p = 1.51e\text{-}6$, *Cohen's d* = 0.70). This effect for VIQ being lower in the ASC 'Impaired' subgroup is somewhat expected given that some portion of variability in RMET performance is known to be linked with VIQ variability [70]. It is also noteworthy that while these effects appear to be consistent with ideas regarding effects of VIQ on RMET performance, they are not strong enough to pass correction for multiple comparisons for the full number of tests run across the entire study. Therefore, we would interpret this result as of some importance, given its links to prior ideas about the effects of VIQ on RMET performance, but it is unlikely that these effects are strong enough to completely explain the presence of these subgroups or the massive reductions observed in RMET total scores. Congruent with these ideas, when we covaried out variability in VIQ in our analyses of RMET total scores, we still found robust evidence for decreased RMET scores in the ASC Impaired group compared to the TD group, and this suggests that while a correlation between VIQ and RMET performance is likely to be real, it does not explain the presence of ASC RMET subgroups identified here. See Supplementary Figure 3.

Age across ASC subgroups

We found no consistent effect for any differences in age between ASC subgroups or TD across both datasets (Discovery: $F(2,691) = 4.49$, $p = 0.01$; 'Impaired' mean = 39.68, 'Unimpaired' mean = 36.94, TD mean = 35.07; 'Impaired' vs 'Unimpaired': $t(376) = 1.64$, $p = 0.10$, *Cohen's d* = 0.23; 'Impaired' vs TD: $t(372) = 2.64$, $p = 0.008$, *Cohen's d* = 0.37; 'Unimpaired' vs TD: $t(634) = 1.99$, $p = 0.04$, *Cohen's d* = 0.15; Replication: $F(2,246) = 1.68$, $p = 0.18$; 'Impaired' mean = 28.14, 'Unimpaired' mean = 26.14, TD mean = 27.84; 'Impaired' vs

'Unimpaired': *t(121)* = 1.35, *p* = 0.17, *Cohen's d* = 0.26; 'Impaired' vs TD: *t(164)* = 0.22, *p* = 0.82, *Cohen's d* = 0.04; 'Unimpaired' vs TD: *t(207)* = -1.73, *p* = 0.08, *Cohen's d* = 0.24). See Supplementary Figure 4.

Autistic traits and trait empathy across ASC subgroups

With regard to autistic traits measured by the AQ, we found no evidence of a consistent difference in autistic traits between ASC subgroups across both datasets, despite the fact that in comparison to TD, there was the known large difference in autistic traits (Discovery: *F(2,691)* = 467.15, *p* = 4.58e-129; 'Impaired' mean = 39.48, 'Unimpaired' mean = 34.96, TD mean = 16.25; 'Impaired' vs 'Unimpaired': *t(376)* = 3.21, *p* = 0.001, *Cohen's d* = 0.45; 'Impaired' vs TD: *t(372)* = 24.35, *p* = 5.12e-79, *Cohen's d* = 3.47; 'Unimpaired' vs TD: *t(634)* = 28.19, *p* = 5.41e-114, *Cohen's d* = 2.23; Replication: *F(2,220)* = 145.56, *p* = 5.35e-41; 'Impaired' mean = 32.21, 'Unimpaired' mean = 30.89, TD mean = 13.88; 'Impaired' vs 'Unimpaired': *t(107)* = 0.70, *p* = 0.48, *Cohen's d* = 0.14; 'Impaired' vs TD: *t(145)* = 13.33, *p* = 5.53e-27, *Cohen's d* = 2.61; 'Unimpaired' vs TD: *t(188)* = 15.76, *p* = 3.14e-36, *Cohen's d* = 2.33). See Supplementary Figure 5.

Similarly, for trait empathy measured by the EQ, there were no consistent differences between ASC subgroups across datasets despite there being a large difference when compared to the TD group (Discovery: *F(2,691)* = 271.90, *p* = 7.78e-88; 'Impaired' mean = 15.79, 'Unimpaired' mean = 23.23, TD mean = 46.12; 'Impaired' vs 'Unimpaired': *t(376)* = -3.55, *p* = 4.23e-4, *Cohen's d* = 0.50; 'Impaired' vs TD: *t(372)* = -17.18, *p* = 3.91e-49, *Cohen's d* = 2.45; 'Unimpaired' vs TD: *t(634)* = -20.81, *p* = 9.75e-74, *Cohen's d* = 1.65; Replication: *F(2,219)* = 100.58, *p* = 1.03e-31; 'Impaired' mean = 22.69, 'Unimpaired' mean = 24.01, TD mean = 46.72; 'Impaired' vs 'Unimpaired': *t(107)* = -0.53, *p* = 0.59, *Cohen's d* = 0.11; 'Impaired' vs TD: *t(144)* = -9.49, *p* = 6.54e-17, *Cohen's d* = 1.87; 'Unimpaired' vs TD: *t(187)* = -12.87, *p* = 1.38e-27, *Cohen's d* = 1.90). See Supplementary Figure 6.

Depression and Anxiety across ASC subgroups

Regarding the severity of depressive symptoms as measured by the BDI within the AIMS dataset, we found no evidence for differences between the ASC subgroups, despite both these subgroups showing elevated scores compared to the TD group (*F(2,214)* = 21.12, *p* = 4.25e-9; 'Impaired' mean = 13.92, 'Unimpaired' mean = 12.43, TD mean = 5.80; 'Impaired' vs 'Unimpaired': *t(119)* = 0.78, *p* = 0.43, *Cohen's d* = 0.15; 'Impaired' vs TD: *t(132)* = 6.27, *p* = 4.67e-9, *Cohen's d* = 1.23; 'Unimpaired' vs TD: *t(177)* = 5.66, *p* = 5.79e-8, *Cohen's d* = 0.88). See Supplementary Figure 7A.

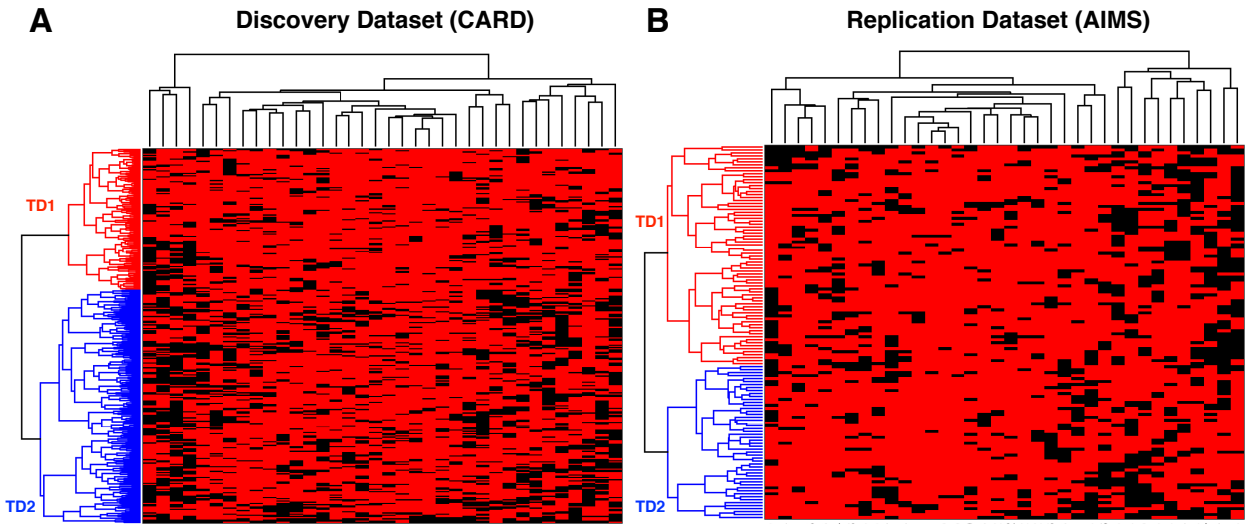
Similarly for the severity of anxiety symptoms measured by the BAI, we also found no evidence of ASC subgroup differentiation, despite both being elevated compared to TD ($F(2,243)$ = 32.23, p = 3.83e-13; 'Impaired' mean = 16.74, 'Unimpaired' mean = 12.74, TD mean = 4.93; 'Impaired' vs 'Unimpaired': $t(119)$ = 1.72, p = 0.08, *Cohen's d* = 0.33; 'Impaired' vs TD: $t(162)$ = 7.96, p = 2.78e-13, *Cohen's d* = 1.45; 'Unimpaired' vs TD: $t(205)$ = 6.62, p = 3.06e-10, *Cohen's d* = 0.94). See Supplementary Figure 7B.

Autism Symptom Severity across ASC subgroups

Finally, with regards to early developmental autism symptom severity as measured by algorithm scores on the ADI-R, we found no evidence of differentiation of ASC Impaired compared to ASC Unimpaired subgroups on the social ($t(97)$ = 1.04, p = 0.29, *Cohen's d* = 0.22), communication ($t(97)$ = 0.51, p = 0.61, *Cohen's d* = 0.10) or repetitive and restricted behavior subscales ($t(97)$ = -0.76, p = 0.44, *Cohen's d* = 0.16). See Supplementary Figure 8.

However, for current symptom severity measured by ADOS algorithm scores, we found that the ASD Impaired subgroup was slightly elevated on both the communication ($t(118)$ = 2.0004, p = 0.04, *Cohen's d* = 0.38), and social interaction subscales ($t(97)$ = 2.34, p = 0.02, *Cohen's d* = 0.45). However, because these effects are somewhat weak, and would not pass correction for multiple testing, we would err on the side of caution when interpreting it as a potential true effect. See Supplementary Figure 9.

Supplementary Figures



Supplementary Figure 1: *This figure shows TD subgroups (red and blue) on clustergrams of both the discovery (A) and replication (B) datasets.*

A **Discovery Dataset (CARD)**

	Impaired	Unimpaired
Male	22 (25.93)	147 (143.06)
Female	36 (32.06)	173 (176.93)

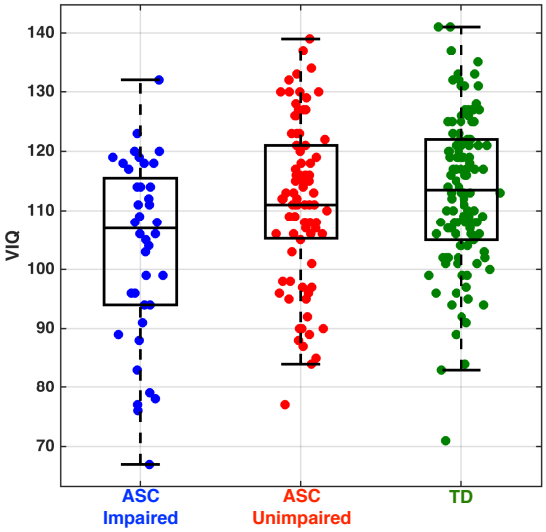
$\chi^2 = 1.27$, $p = 0.25$

B **Replication Dataset (AIMS)**

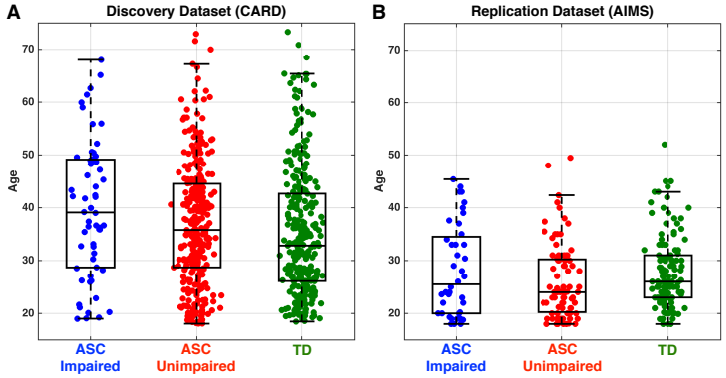
	Impaired	Unimpaired
Male	26 (27.64)	59 (57.35)
Female	14 (12.35)	24 (25.64)

$\chi^2 = 0.46$, $p = 0.49$

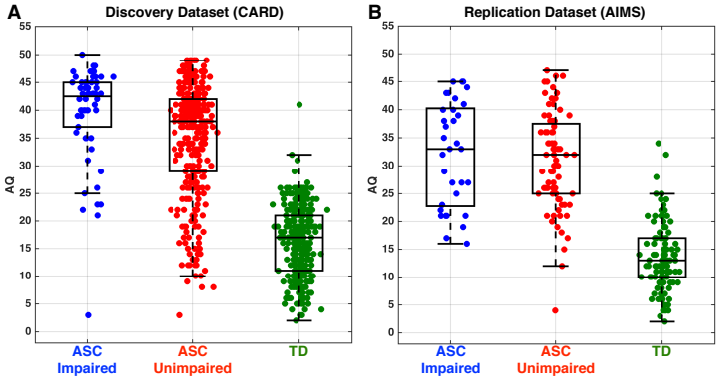
Supplementary Figure 2: *This figure shows contingency tables with counts of the number of individuals in both datasets (A, Discovery CARD dataset; B, Replication AIMS dataset) that fall into the different ASC subgroups and who are either male or female. The numbers within each cell correspond to the actual count and the expected count within the parentheses.*



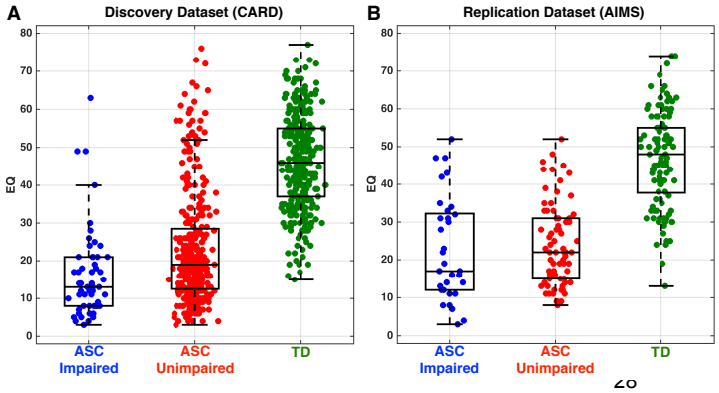
Supplementary Figure 3: *This figure shows the scatter-boxplot of verbal IQ (VIQ) measured within the replication dataset (AIMS) across all groups.*



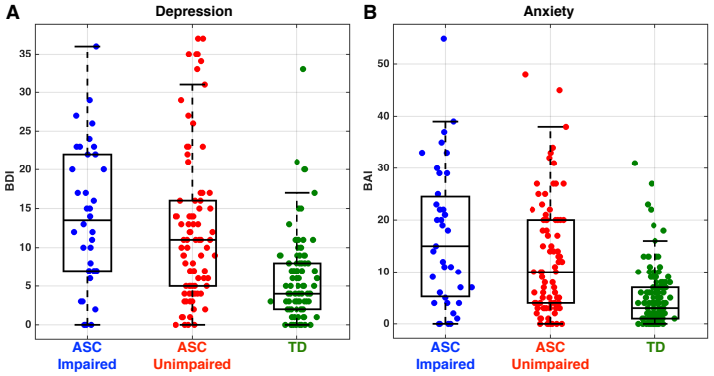
Supplementary Figure 4: *This figure shows the scatter-boxplots across both datasets (A, Discovery CARD dataset; B, Replication AIMS dataset) for age across all groups.*



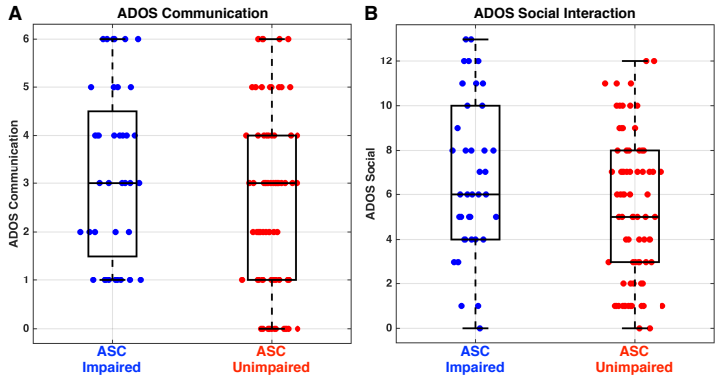
Supplementary Figure 5: *This figure shows the scatter-boxplots across both datasets (A, Discovery CARD dataset; B, Replication AIMS dataset) for autistic traits measured by the AQ across all groups.*



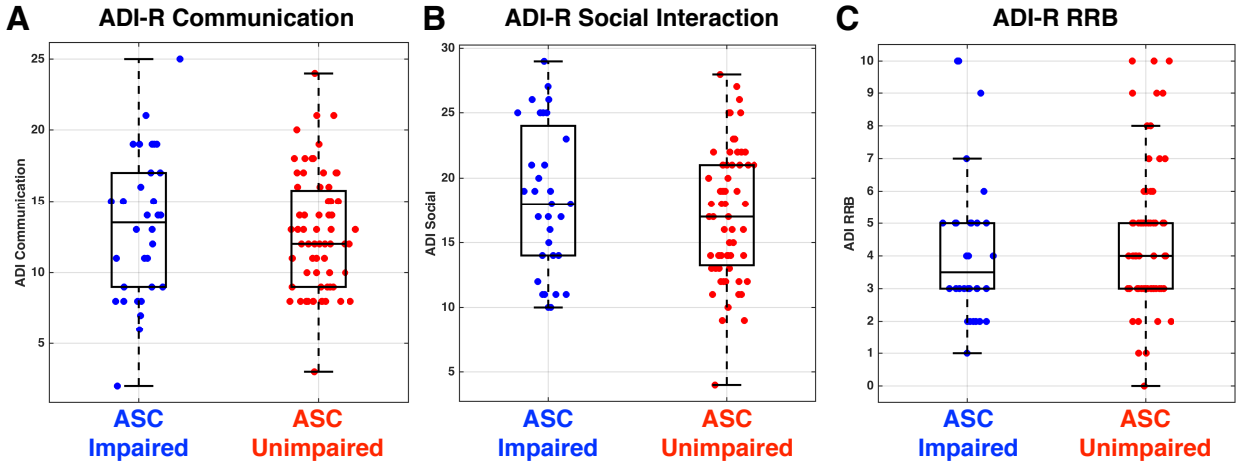
Supplementary Figure 6: *This figure shows the scatter-boxplots across both datasets (A, Discovery CARD dataset; B, Replication AIMS dataset) for trait* empathy measured by the EQ across all groups.



Supplementary Figure 7: *This figure shows the scatter-boxplots of depression (A) or anxiety (B) symptom scores measured by the BDI and BAI respectively for all groups. This data was measured only within the replication dataset (AIMS).*



Supplementary Figure 8: *This figure shows the scatter-boxplots of ADOS communication (A) and social interaction (B) algorithm scores for the ASC subgroups. This data was measured only within the replication dataset (AIMS).*



Supplementary Figure 9: *This figure shows the scatter-boxplots of ADI-R communication (A), social interaction (B), or restricted/repetitive behavior (C) algorithm scores for the ASC subgroups. This data was measured only within the replication dataset (AIMS).*