Current Biology, Volume *29*

Supplemental Information

Deconstructing Theory-of-Mind Impairment in High-Functioning Adults with Autism

Isabelle A. Rosenthal, Cendri A. Hutcherson, Ralph Adolphs, and Damian A. Stanley

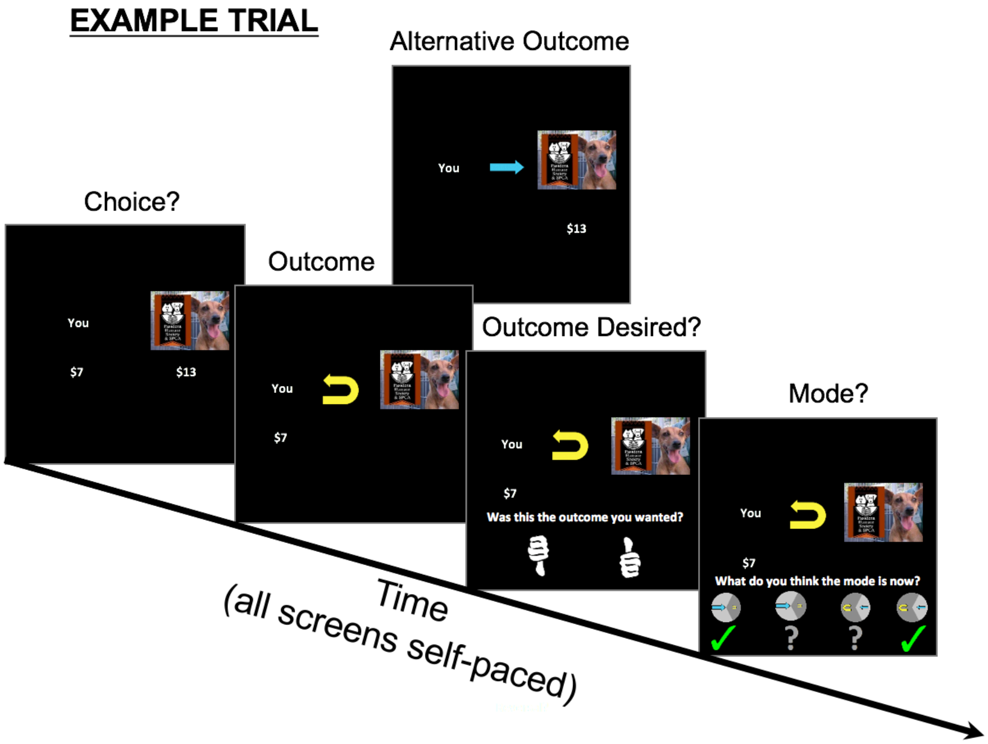
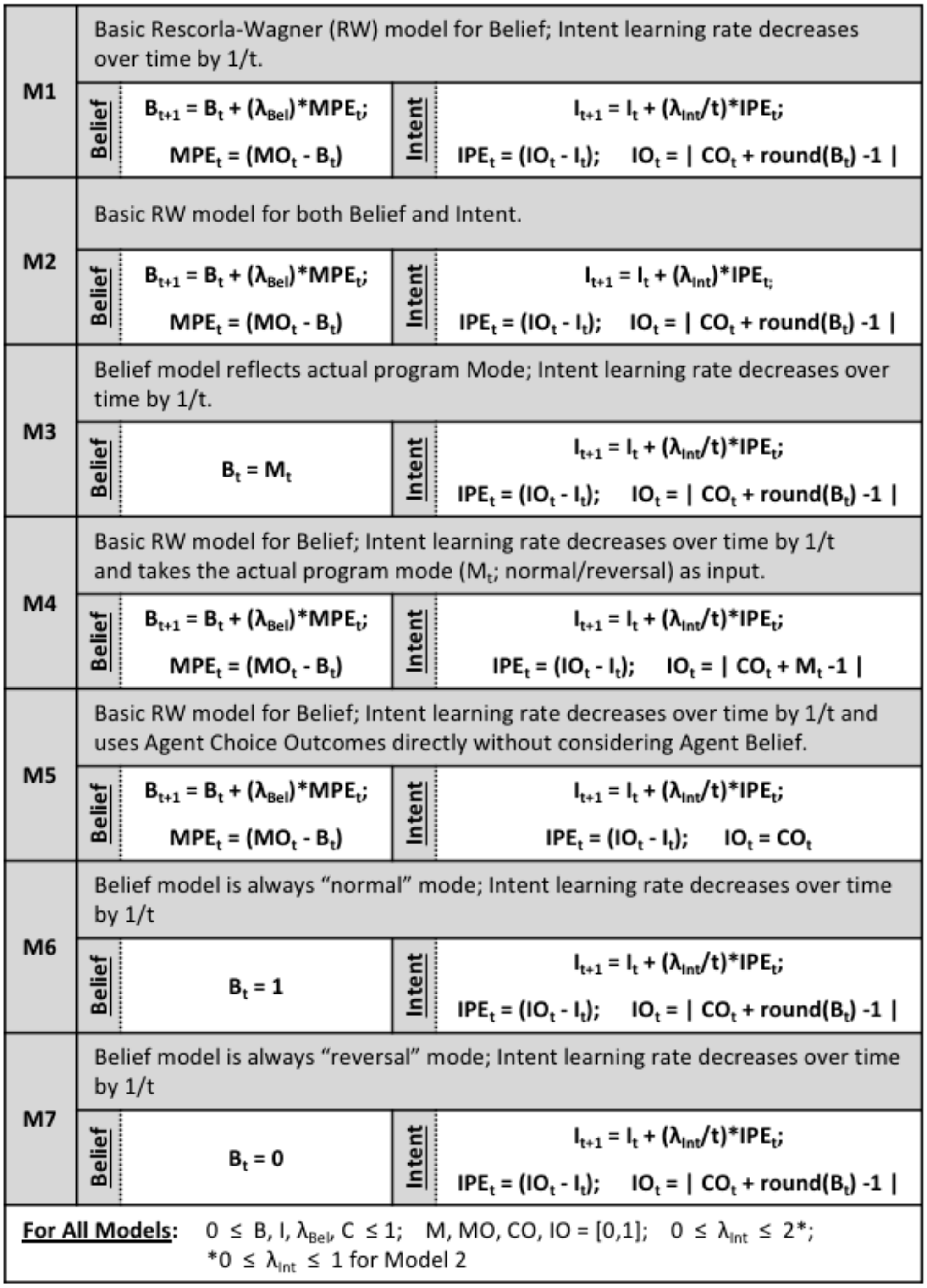


Figure S1. Schematic of Charity Task. Related to Figure 1, Task Performance Accuracy, and STAR Methods.

Participants and Agents performed a charitable giving task in which they decided whether to give money to one of three charities (donate), or take it for themselves (take). On each trial, the participant (or Agent) was shown a picture of one of three charities on one side of the screen, with "$10" displayed underneath it, and the word "you" displayed on the other side of the screen, with a dollar amount (ranging from $7-$13) displayed underneath it. The participant then made a choice (donate or take). The computer program had two modes (or contexts), "normal" and "reversal". In "normal" mode, the program intervened and reversed the participants' choices on 36% of trials (leaving them unaltered on 64% of trials). In "reversal" mode the opposite was true, the program reversed the participants' choices on 64% of trials (leaving choices unaltered on 36% of trials). Therefore, to obtain their desired outcomes most of the time, participants would have to reverse their decisions (i.e. choose what they didn't want) when the program was in "reversal" mode. Importantly, participants were not explicitly aware of the current program mode (except during practice; see below), and instead needed to track it by observing how often the computer was reversing their decisions. Participants were instructed that the mode was "stable across multiple trials" (in actuality 3-12 trials; see Agent data creation below for details), giving them time to learn, and that every so often it would change. Following the participant's choice, the computer's action was displayed (a blue straight arrow indicated that the choice happened as intended, a yellow curved arrow indicated that the choice had been reversed), and non-chosen dollar amount was removed from the screen. Finally, the participant answered 2 follow-up questions: "Was this the outcome you wanted?" (select thumbs up icon for "yes", thumbs down icon for "no") and "what do you think is the mode now?" Answers were given in the form of 4-alternative-forced-choice across icons representing "Definitely Reversal", "Maybe Reversal", "Maybe Normal", "Definitely Normal". Presentation of all screens was self-paced. The three charities (The Southeast Alaska

Conservation Council, Canine Assistants, Pasadena Humane Society and SPCA) were selected based on a previous charitable giving experiment[S1] which found that they were equally preferred by individuals with ASD and matched controls. To facilitate understanding of the mode structure, participants completed a series of practice trials (21-25) in which they were explicitly told the program mode ("normal" or "reversal") via the display of a mode-specific icon.

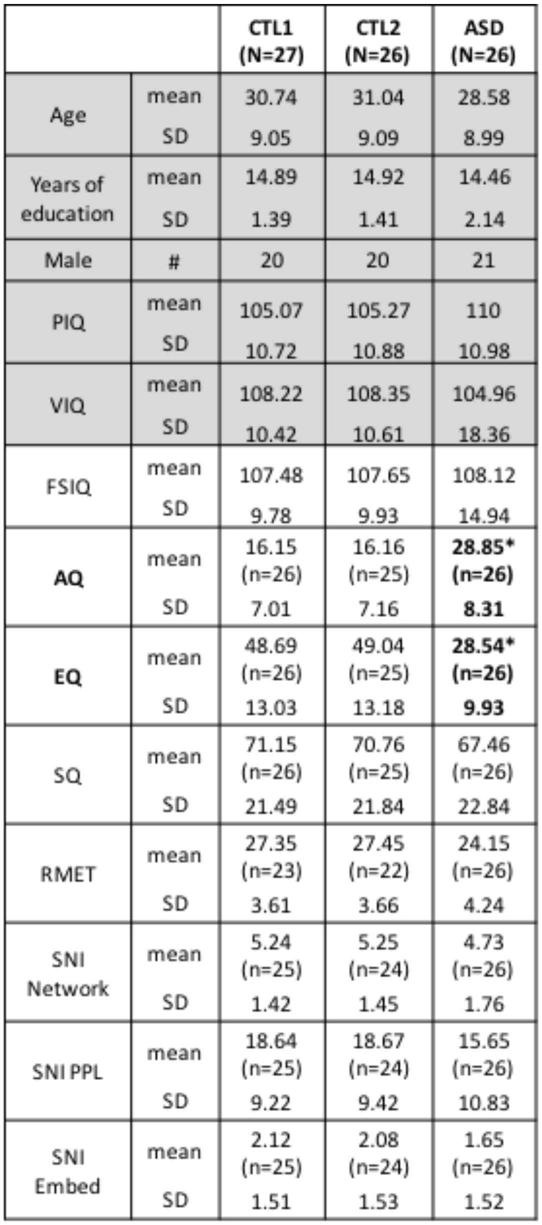


Model	Description	Belief	Intent
M1	Basic Rescorla-Wagner (RW) model for Belief; Intent learning rate decreases over time by 1/t.	$B_{t+1} = B_t + (\lambda_{Bel})*MPE_t$; $MPE_t = (MO_t - B_t)$	$I_{t+1} = I_t + (\lambda_{Int}/t)*IPE_t$; $IPE_t = (IO_t - I_t)$; $IO_t = \vert CO_t + round(B_t) -1 \vert$
M2	Basic RW model for both Belief and Intent.	$B_{t+1} = B_t + (\lambda_{Bel})*MPE_t$; $MPE_t = (MO_t - B_t)$	$I_{t+1} = I_t + (\lambda_{Int})*IPE_{t;}$ $IPE_t = (IO_t - I_t)$; $IO_t = \vert CO_t + round(B_t) -1 \vert$
M3	Belief model reflects actual program Mode; Intent learning rate decreases over time by 1/t.	$B_t = M_t$	$I_{t+1} = I_t + (\lambda_{Int}/t)*IPE_t$; $IPE_t = (IO_t - I_t)$; $IO_t = \vert CO_t + round(B_t) -1 \vert$
M4	Basic RW model for Belief; Intent learning rate decreases over time by 1/t and takes the actual program mode (M_t; normal/reversal) as input.	$B_{t+1} = B_t + (\lambda_{Bel})*MPE_t$; $MPE_t = (MO_t - B_t)$	$I_{t+1} = I_t + (\lambda_{Int}/t)*IPE_t$; $IPE_t = (IO_t - I_t)$; $IO_t = \vert CO_t + M_t -1 \vert$
M5	Basic RW model for Belief; Intent learning rate decreases over time by 1/t and uses Agent Choice Outcomes directly without considering Agent Belief.	$B_{t+1} = B_t + (\lambda_{Bel})*MPE_t$; $MPE_t = (MO_t - B_t)$	$I_{t+1} = I_t + (\lambda_{Int}/t)*IPE_t$; $IPE_t = (IO_t - I_t)$; $IO_t = CO_t$
M6	Belief model is always "normal" mode; Intent learning rate decreases over time by 1/t	$B_t = 1$	$I_{t+1} = I_t + (\lambda_{Int}/t)*IPE_t$; $IPE_t = (IO_t - I_t)$; $IO_t = \vert CO_t + round(B_t) -1 \vert$
M7	Belief model is always "reversal" mode; Intent learning rate decreases over time by 1/t	$B_t = 0$	$I_{t+1} = I_t + (\lambda_{Int}/t)*IPE_t$; $IPE_t = (IO_t - I_t)$; $IO_t = \vert CO_t + round(B_t) -1 \vert$

For All Models: $0 \le B, I, \lambda_{Bel}, C \le 1$; $M, MO, CO, IO = [0,1]$; $0 \le \lambda_{Int} \le 2$*;
*$0 \le \lambda_{Int} \le 1$ for Model 2

Figure S2. Equations for Theory of Mind models. Related to Figures 3, 4, and STAR Methods.

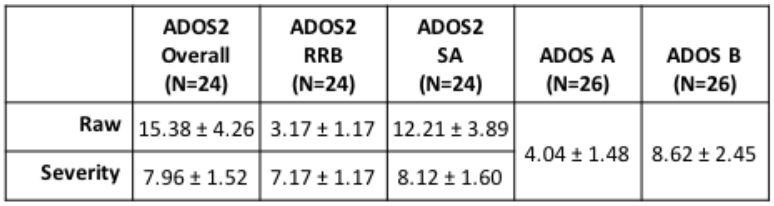
B_t and I_t are the Mentalizer's probabilistic estimate of the Agent's belief about the program mode (M_t; "Normal" or "Reversal") and the Agent's intention to donate to the charity, respectively, on trial t. λ_{Bel} and λ_{Int} are the Belief and Intent learning rates, respectively; MPE_t and MO_t are the Mode Prediction Error and Mode Outcome on trial t. IPE_t and IO_t are the Intent Prediction Error and Intent Outcome on trial t, and CO_t is the Choice Outcome.



		CTL1 (N=27)	CTL2 (N=26)	ASD (N=26)
Age	mean	30.74	31.04	28.58
	SD	9.05	9.09	8.99
Years of education	mean	14.89	14.92	14.46
	SD	1.39	1.41	2.14
Male	#	20	20	21
PIQ	mean	105.07	105.27	110
	SD	10.72	10.88	10.98
VIQ	mean	108.22	108.35	104.96
	SD	10.42	10.61	18.36
FSIQ	mean	107.48	107.65	108.12
	SD	9.78	9.93	14.94
AQ	mean	16.15 (n=26)	16.16 (n=25)	**28.85* (n=26)**
	SD	7.01	7.16	**8.31**
EQ	mean	48.69 (n=26)	49.04 (n=25)	**28.54* (n=26)**
	SD	13.03	13.18	**9.93**
SQ	mean	71.15 (n=26)	70.76 (n=25)	67.46 (n=26)
	SD	21.49	21.84	22.84
RMET	mean	27.35 (n=23)	27.45 (n=22)	24.15 (n=26)
	SD	3.61	3.66	4.24
SNI Network	mean	5.24 (n=25)	5.25 (n=24)	4.73 (n=26)
	SD	1.42	1.45	1.76
SNI PPL	mean	18.64 (n=25)	18.67 (n=24)	15.65 (n=26)
	SD	9.22	9.42	10.83
SNI Embed	mean	2.12 (n=25)	2.08 (n=24)	1.65 (n=26)
	SD	1.51	1.53	1.52

Table S1. Summary of demographic information and survey data. Related to Figures 2-4, Results, and STAR Methods.

Shaded cells indicate measures that were used to match groups. Bold type and *'s indicate measures for which ASD data were significantly different (controlling for multiple comparisons) than control data. Note that there was a significant difference between ASD and CTL for the RMET task (p=0.007) but it did not survive Bonferroni correction.



	ADOS2 Overall (N=24)	ADOS2 RRB (N=24)	ADOS2 SA (N=24)	ADOS A (N=26)	ADOS B (N=26)
Raw	15.38 ± 4.26	3.17 ± 1.17	12.21 ± 3.89	4.04 ± 1.48	8.62 ± 2.45
Severity	7.96 ± 1.52	7.17 ± 1.17	8.12 ± 1.60		

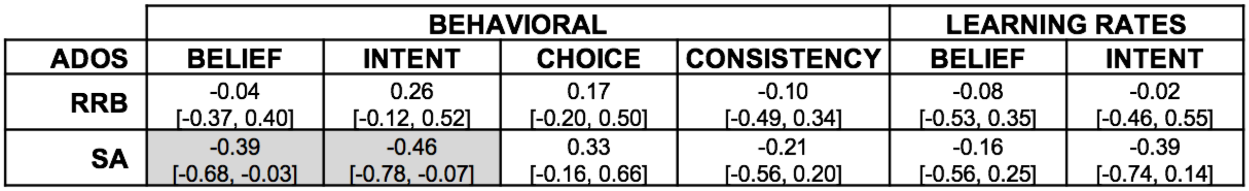
Table S2. Autism Diagnostic Observation Schedule scores. Related to Figures 2-4, and STAR Methods.

Autism Diagnostic Observation Schedule (ADOS)[S2] scores (mean ± standard deviation) for the ASD participants. When available (n=24) ADOS-2[S3] scores are provided (both raw and calibrated severity scores[S4]).

Group	Overall Accuracy				Learning: (Accuracy$_{Last30}$ –Accuracy$_{First30}$)	
	Belief	Intent	Choice	Consistency	Intent	Choice
CTL1	0.65 [0.60, 0.69]	0.66 [0.59, 0.73]	0.57 [0.55, 0.61]	0.74 [0.66, 0.81]	0.14 [0.07, 0.22]	0.01 [-0.04, 0.05]
CTL2	0.66 [0.62, 0.71]	0.61 [0.54, 0.70]	0.55 [0.52, 0.59]	0.76 [0.69, 0.83]	0.12 [0.06, 0.20]	0.01 [-0.04, 0.09]
ASD	0.59 [0.56, 0.63]	0.56 [<0.50, 0.63]	0.51 [0.49, 0.52]	0.70 [0.63, 0.76]	0.02 [-0.05, 0.08]	0.00 [-0.06, 0.06]

Table S3. Mentalizer Task Behavioral Performance. Related to Figure 2, Task Performance Accuracy, and STAR Methods.

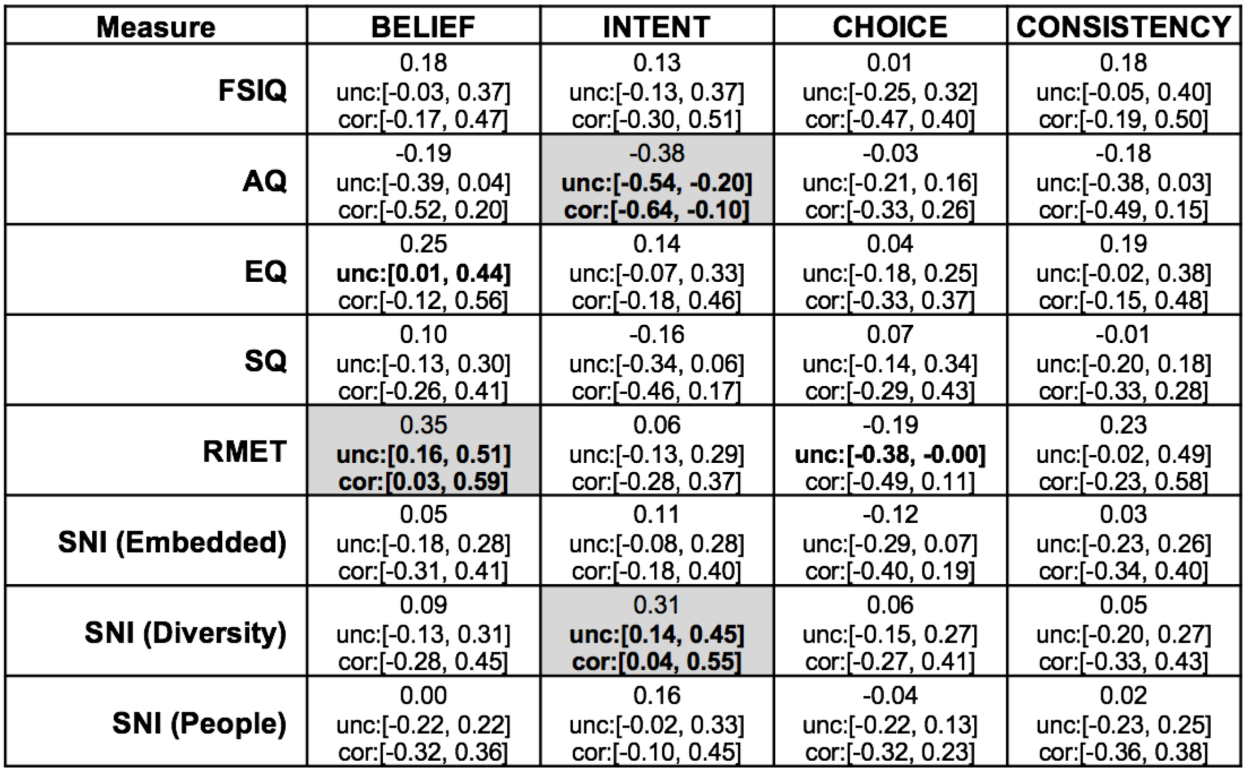
Table of mentalizer task performance: Overall Accuracy for Belief, Intent, Choice, and Consistency, as well as Learning (change in mean accuracy from the first 30 to the last 30 trials) for Intent and Choice. Shaded boxes indicate those means for which bootstrapped 95% confidence intervals excluded chance accuracy (0.5) and chance learning (0.0) performance. As mentioned in the main text, CTL performance was above chance for all measures except Choice Learning, while ASD performance was specifically impaired for Intent and Choice Overall Accuracy, as well as Intent Learning (Accuracy$_{Last30}$-Accuracy$_{First30}$). This pattern of results was robust to the number of trials used to calculate learning (e.g. 20 or 40). Note: Baseline Intent performance (average of first 10 trials) did not differ from chance in any group.



	BEHAVIORAL				LEARNING RATES	
ADOS	BELIEF	INTENT	CHOICE	CONSISTENCY	BELIEF	INTENT
RRB	-0.04 [-0.37, 0.40]	0.26 [-0.12, 0.52]	0.17 [-0.20, 0.50]	-0.10 [-0.49, 0.34]	-0.08 [-0.53, 0.35]	-0.02 [-0.46, 0.55]
SA	-0.39 [-0.68, -0.03]	-0.46 [-0.78, -0.07]	0.33 [-0.16, 0.66]	-0.21 [-0.56, 0.20]	-0.16 [-0.56, 0.25]	-0.39 [-0.74, 0.14]

Table S4. Correlations between mentalizer task performance and ADOS symptom severity. Related to Figures 2, 3, and Analysis of Individual Differences.

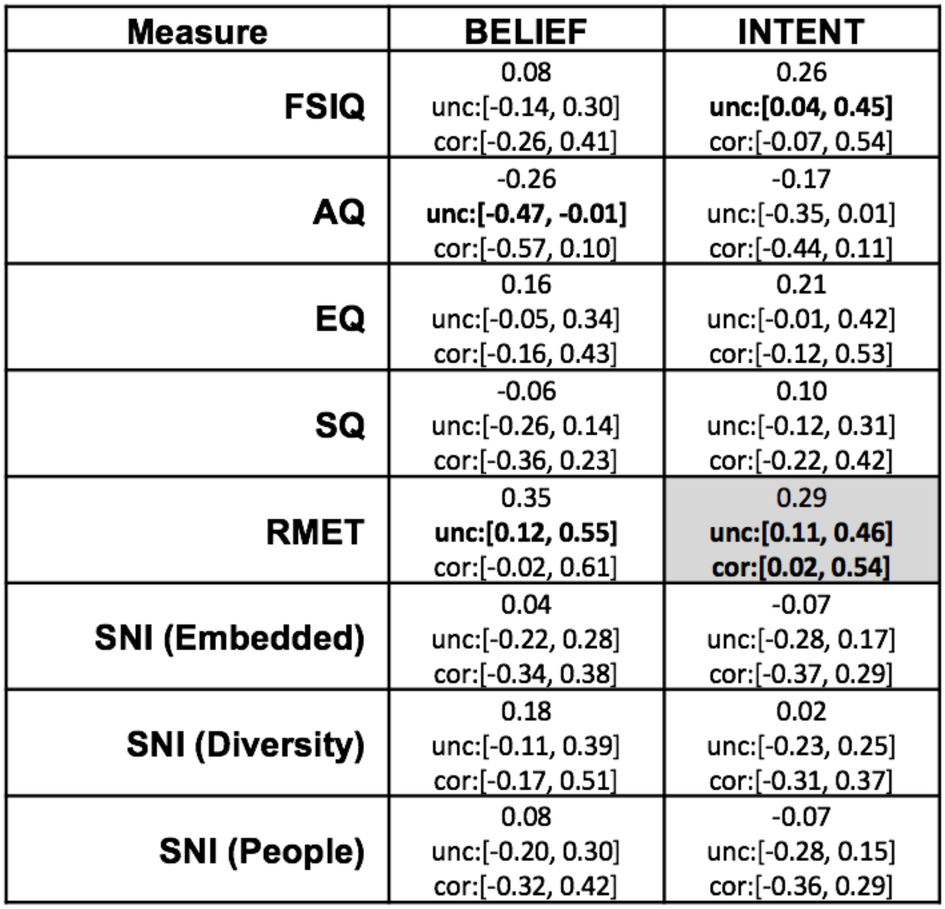
Correlation coefficients (Pearson's r with 95% confidence intervals) describing the relationships between ADOS Severity Scores[S4] (Restricted and Repetitive Behaviors [RRB] and Social Affect [SA]) and task performance (overall accuracy for belief and consistency, change in accuracy for intent and choice, as well as belief and intent learning rates from model fitting). Shaded boxes indicate 95% confidence intervals that excluded zero. Only SA scores exhibited a significant relationship with task performance (Belief and Intent).



Measure	BELIEF	INTENT	CHOICE	CONSISTENCY
FSIQ	0.18 unc:[-0.03, 0.37] cor:[-0.17, 0.47]	0.13 unc:[-0.13, 0.37] cor:[-0.30, 0.51]	0.01 unc:[-0.25, 0.32] cor:[-0.47, 0.40]	0.18 unc:[-0.05, 0.40] cor:[-0.19, 0.50]
AQ	-0.19 unc:[-0.39, 0.04] cor:[-0.52, 0.20]	-0.38 **unc:[-0.54, -0.20]** **cor:[-0.64, -0.10]**	-0.03 unc:[-0.21, 0.16] cor:[-0.33, 0.26]	-0.18 unc:[-0.38, 0.03] cor:[-0.49, 0.15]
EQ	0.25 **unc:[0.01, 0.44]** cor:[-0.12, 0.56]	0.14 unc:[-0.07, 0.33] cor:[-0.18, 0.46]	0.04 unc:[-0.18, 0.25] cor:[-0.33, 0.37]	0.19 unc:[-0.02, 0.38] cor:[-0.15, 0.48]
SQ	0.10 unc:[-0.13, 0.30] cor:[-0.26, 0.41]	-0.16 unc:[-0.34, 0.06] cor:[-0.46, 0.17]	0.07 unc:[-0.14, 0.34] cor:[-0.29, 0.43]	-0.01 unc:[-0.20, 0.18] cor:[-0.33, 0.28]
RMET	0.35 **unc:[0.16, 0.51]** **cor:[0.03, 0.59]**	0.06 unc:[-0.13, 0.29] cor:[-0.28, 0.37]	-0.19 **unc:[-0.38, -0.00]** cor:[-0.49, 0.11]	0.23 unc:[-0.02, 0.49] cor:[-0.23, 0.58]
SNI (Embedded)	0.05 unc:[-0.18, 0.28] cor:[-0.31, 0.41]	0.11 unc:[-0.08, 0.28] cor:[-0.18, 0.40]	-0.12 unc:[-0.29, 0.07] cor:[-0.40, 0.19]	0.03 unc:[-0.23, 0.26] cor:[-0.34, 0.40]
SNI (Diversity)	0.09 unc:[-0.13, 0.31] cor:[-0.28, 0.45]	0.31 **unc:[0.14, 0.45]** **cor:[0.04, 0.55]**	0.06 unc:[-0.15, 0.27] cor:[-0.27, 0.41]	0.05 unc:[-0.20, 0.27] cor:[-0.33, 0.43]
SNI (People)	0.00 unc:[-0.22, 0.22] cor:[-0.32, 0.36]	0.16 unc:[-0.02, 0.33] cor:[-0.10, 0.45]	-0.04 unc:[-0.22, 0.13] cor:[-0.32, 0.23]	0.02 unc:[-0.23, 0.25] cor:[-0.36, 0.38]

Table S5. Correlations between mentalizer task behavioral performance and survey measures. Related to Figure 2 and Analysis of Individual Differences.

Correlation coefficients (Pearson's r) describing the relationships between performance (overall accuracy for belief and consistency and change in accuracy for intent and choice) and other measures. For the purposes of these analyses, data were collapsed across groups. Both uncorrected (unc) and Bonferroni corrected (cor) 95% confidence intervals are presented in brackets under the Pearson's r value. Bolded text indicates 95% confidence intervals that excluded zero, shaded boxes indicate those correlations for which corrected bootstrapped 95% confidence intervals excluded zero.



Measure	BELIEF	INTENT
FSIQ	0.08 unc:[-0.14, 0.30] cor:[-0.26, 0.41]	0.26 **unc:[0.04, 0.45]** cor:[-0.07, 0.54]
AQ	-0.26 **unc:[-0.47, -0.01]** cor:[-0.57, 0.10]	-0.17 unc:[-0.35, 0.01] cor:[-0.44, 0.11]
EQ	0.16 unc:[-0.05, 0.34] cor:[-0.16, 0.43]	0.21 unc:[-0.01, 0.42] cor:[-0.12, 0.53]
SQ	-0.06 unc:[-0.26, 0.14] cor:[-0.36, 0.23]	0.10 unc:[-0.12, 0.31] cor:[-0.22, 0.42]
RMET	0.35 **unc:[0.12, 0.55]** cor:[-0.02, 0.61]	▨ 0.29 **unc:[0.11, 0.46]** **cor:[0.02, 0.54]**
SNI (Embedded)	0.04 unc:[-0.22, 0.28] cor:[-0.34, 0.38]	-0.07 unc:[-0.28, 0.17] cor:[-0.37, 0.29]
SNI (Diversity)	0.18 unc:[-0.11, 0.39] cor:[-0.17, 0.51]	0.02 unc:[-0.23, 0.25] cor:[-0.31, 0.37]
SNI (People)	0.08 unc:[-0.20, 0.30] cor:[-0.32, 0.42]	-0.07 unc:[-0.28, 0.15] cor:[-0.36, 0.29]

Table S6. Correlations between mentalizer task model parameters and survey measures. Related to Figure 3 and Analysis of Individual Differences.

Correlation coefficients (Pearson's r) describing the relationships between estimated model parameters (belief and intent learning rate) from the individual-level fits and survey measures. Both uncorrected (unc) and Bonferroni corrected (cor) 95% confidence intervals are presented in brackets under the Pearson's r value. Bolded text indicates 95% confidence intervals that excluded zero, shaded boxes indicate those correlations for which corrected bootstrapped 95% confidence intervals excluded zero. To ensure no bias in correlational analyses involving model parameters, all data from the mentalizer task (CTL1, CTL2, and ASD) were combined and fit with a single hierarchical model (using the *a priori* model M1). Our reasoning was that, given that model 1 best describes control performance (and therefore putatively intact ToM processing), the extent to which it captures individual participant performance across all groups might correlate with other measures that have been used to assess autism-like traits and social ability in the literature.

Supplemental References:

1. Lin, A., Rangel, A., and Adolphs, R. (2012). Impaired Learning of Social Compared to Monetary Rewards in Autism. Front Neurosci *6*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3461406/ [Accessed November 20, 2018].

2. Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Leventhal, B.L., DiLavore, P.C., Pickles, A., and Rutter, M. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord *30*, 205–223.

3. Lord, C., Rutter, M., DiLavore, P.C., Risi, S., Gotham, K., and Bishop, S.L. (2012). ADOS-2: Autism Diagnostic Observation Schedule, second edition. Part 1: Modules 1–4.

4. Hus, V., and Lord, C. (2014). The autism diagnostic observation schedule, module 4: revised algorithm and standardized severity scores. J Autism Dev Disord *44*, 1996–2012.