Responsive Robots

Preliminary Analysis of Trust and Performance Data

Shauna Heron

2025-12-01

⚠ Warning

These are preliminary results and analyses. Please do not distribute or cite without permission of the authors.

Demographics and Baseline Characteristics

The two groups look very similar on gender, age band, program of study, prior robot experience, NARS, and Need for Cognition; none of these baseline differences are statistically significant; suggesting post-test differences are unlikely to be driven by obvious demographic or attitudinal imbalances.

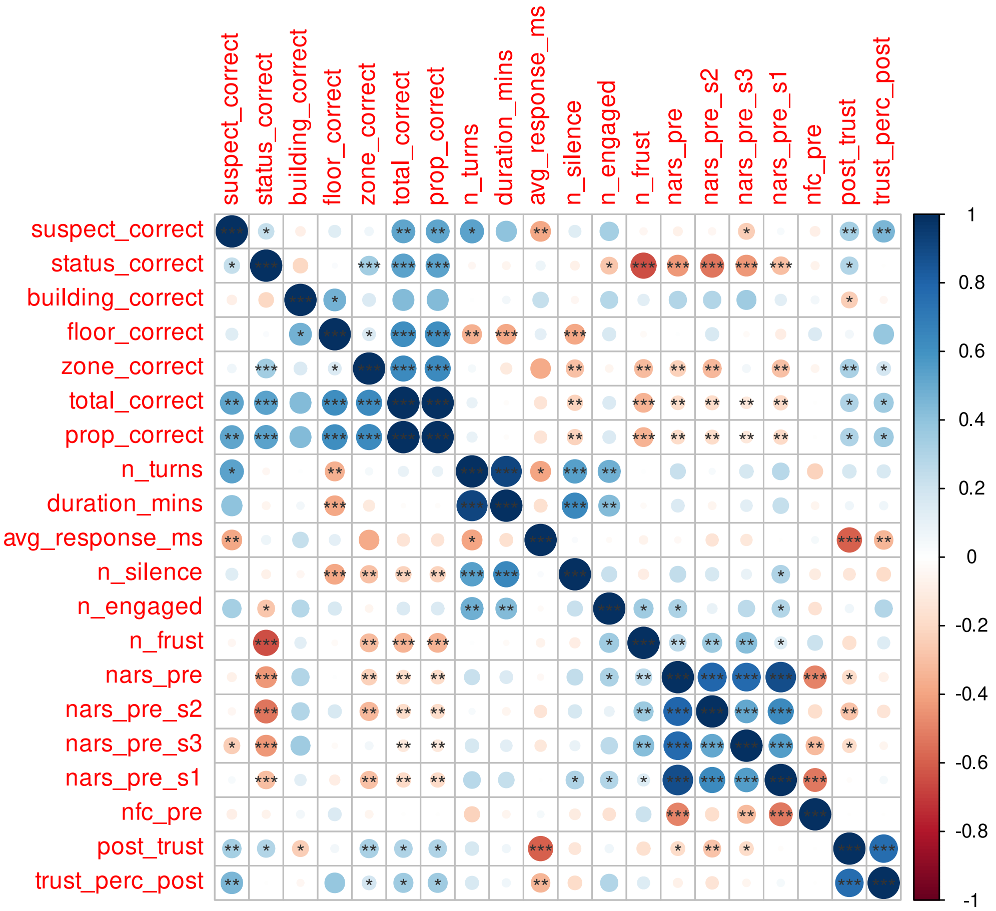
Characteristic	N	CONTROL N = 8	EXPERIMENTAL N = 13	p-value[1]
Gender, n / N (%)	19			>0.99
Woman		3 / 7 (43%)	5 / 12 (42%)	
Man		4 / 7 (57%)	7 / 12 (58%)	
Missing		1	1	
Age Group, n / N (%)	19			0.48
18-24		2 / 7 (29%)	7 / 12 (58%)	
25-34		3 / 7 (43%)	3 / 12 (25%)	
34-44		2 / 7 (29%)	2 / 12 (17%)	
Missing		1	1	
Program, n / N (%)	19			0.22
Psychology		0 / 7 (0%)	2 / 12 (17%)	
Engineering		1 / 7 (14%)	0 / 12 (0%)	

[1] Fisher's exact test; Wilcoxon rank sum test

Characteristic	**N**	**CONTROL** N = 8	**EXPERIMENTAL** N = 13	**p-value**[1]
Computer Science		6 / 7 (86%)	7 / 12 (58%)	
Other		0 / 7 (0%)	3 / 12 (25%)	
Missing		1	1	
Experience w/ Robots, n / N (%)	21	4 / 8 (50%)	5 / 13 (38%)	0.67
Native English Speaker, n / N (%)	21			0.34
FALSE		7 / 8 (88%)	8 / 13 (62%)	
TRUE		1 / 8 (13%)	5 / 13 (38%)	
NARS, Mean (SD)	21	43 (7)	40 (10)	0.45
nars_pre_s2, Mean (SD)	21	16.88 (3.38)	14.90 (3.62)	0.35
nars_pre_s3, Mean (SD)	21	10.5 (2.7)	9.4 (3.5)	0.56
nars_pre_s1, Mean (SD)	21	15.9 (4.5)	16.4 (5.8)	0.69
Need for Cognition, Mean (SD)	21	3.21 (0.78)	3.35 (1.06)	0.97

[1] Fisher's exact test; Wilcoxon rank sum test

Correlation Matrix of Key Variables



Post-test robot trust ratings

On the overall post-interaction trust score, the responsive-Misty group reports substantially higher trust than the control group (M ≈ 78 vs 57; medium–large effect size, d ≈ 0.67).

In the mixed-effects models (random intercepts for participant and item, controlling for NARS and NFC), group remains a significant predictor of higher trust ratings even after we add native-English status and prior robot experience. Estimated group effect is around +26 trust points (on a scale of 0 to 100) for the responsive condition.

Unexpectedly, being a native English speaker is also associated with higher trust ratings (≈ +21 points), and there are a few specific trust items (reliable, responsive, trustworthy) where the group × item interaction reaches significance, but the general pattern is that the responsive robot is trusted more across items.

Task accuracy

For task *performance* (accuracy on the mystery task items), the biggest predictor isn't group, it's language: native English speakers have much higher odds of determing the correct answer (OR ≈ 8.2, 95% CI ~2–34).

The treatment (responsive vs control) goes in the expected direction (OR < 1 for control) but isn't statistically significant in this small sample. This suggests that speech recognition / accent issues and general language comprehension are strongly constraining task success, and may be masking any performance advantage of the responsive robot. On the other hand, both robots were using the same ASR and dialogue system with access to the same information, so we shouldn't necessarily see an effect here.

Interpretation so far

Trust: despite the small sample, we already see a pretty robust treatment effect on perceived trust in the robot that holds up in the multilevel models after controlling for attitudes and language background.

Performance: accuracy is dominated by language / accent effects, which we hadn't explicitly designed for. This may be a useful finding in itself for HRI and ASR-based systems, but it complicates clean inferences about responsiveness ⟶ task performance. Though we do see a trend in a positive direction.

My read is that these data are very encouraging as a pilot! We have evidence that the responsive, proactive behaviour of the responsive manipulation was doing something psychologically meaningful (higher trust), even with only 21 participants and an imbalanced design (where it should be difficult to detect small effects!). We've also uncovered an important design issue: language background and STT quality are non-trivial confounds for both trust and performance, especially in an international / CS student population.

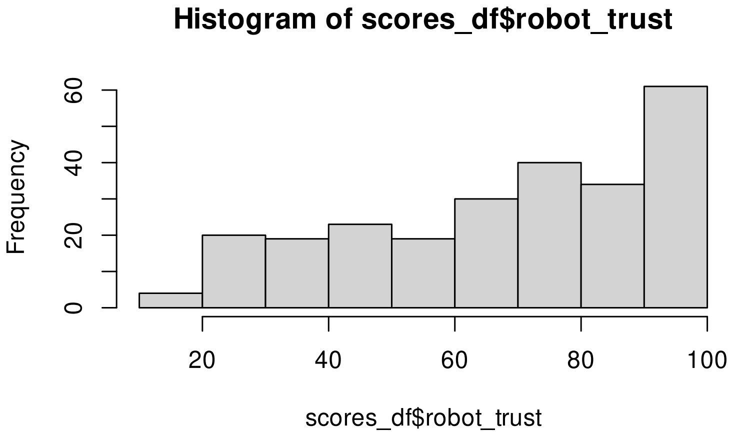
In terms of what to do now:

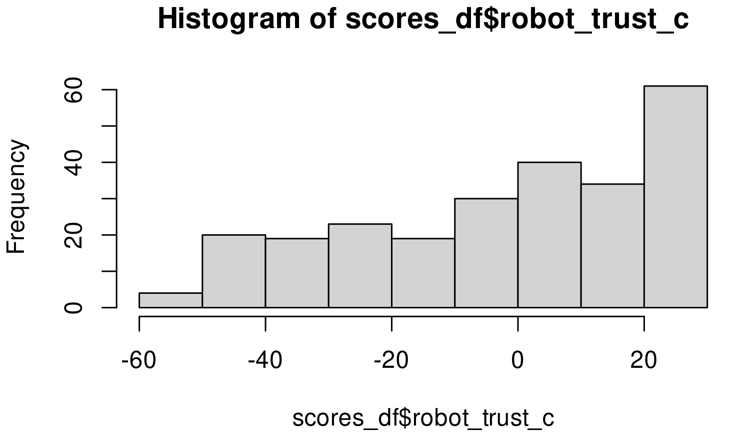
If we can recruit a few more participants without too much pain, I'd love to top up the control group and try to get a better balance of native vs non-native speakers in each condition. This may not be possible in this current study iteration but maybe could expand to a larger thesis?? :) Even getting to something like ~12–15 per group would help the trust analyses and give us a bit more power on accuracy (hopefully our effect wouldn't disappear :D!).

If we can't swing more recruitment, is it defensible to freeze the sample and explicitly frame this as a pilot study?: "trust effect looks promising; performance is constrained by language/ASR issues; next study will (a) stratify or screen on language background, and (b) adjust the dialogue / ASR pipeline to be more robust."

Either way, these results already give us a solid story: responsive Misty is perceived as more trustworthy, and the study has highlighted the need to take language and accent into account when designing HRI tasks that rely on spoken interaction.

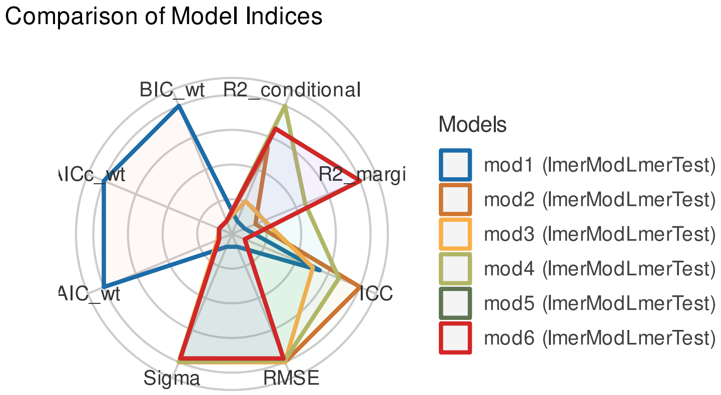
Take a look at some of the effect plots I added below:





```
Data: scores_df
Models:
mod1: robot_trust ~ group + nars_pre + nfc_pre + (1 | session_id) + (1 | trust_items)
mod3: robot_trust ~ group * trust_items + nars_pre + nfc_pre + (1 | session_id)
mod2: robot_trust ~ group * trust_items + nars_pre + nfc_pre + (1 | session_id) + (1 |
trust_items)
mod4: robot_trust ~ group * trust_items + native_english + nars_pre + nfc_pre + (1 |
session_id) + (1 | trust_items)
mod5: robot_trust ~ group * trust_items + native_english + robot_xp + nars_pre +
nfc_pre + (1 | session_id) + (1 | trust_items)
mod6: robot_trust ~ group * trust_items + native_english + robot_xp + nars_pre +
nfc_pre + (1 | session_id) + (1 | trust_items)
     npar    AIC    BIC  logLik -2*log(L)   Chisq Df Pr(>Chisq)
mod1    7 2104.7 2129.3 -1045.3    2090.7
mod3   28 2113.6 2212.2 -1028.8    2057.6 33.1396 21   0.044692 *
mod2   29 2115.6 2217.7 -1028.8    2057.6  0.0000  1   0.999999
mod4   30 2110.5 2216.1 -1025.2    2050.5  7.0923  1   0.007742 **
```

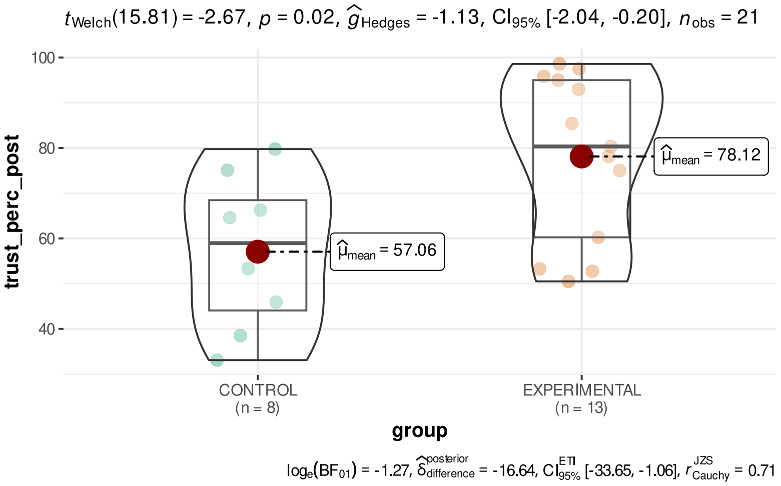
```
mod5   31 2103.7 2212.9 -1020.9    2041.7  8.7208  1   0.003146 **
mod6   31 2103.7 2212.9 -1020.9    2041.7  0.0000  0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
# Comparison of Model Performance Indices

Name |           Model | R2 (cond.) | R2 (marg.) |   ICC |   RMSE |  Sigma
---------------------------------------------------------------------------
mod4 | lmerModLmerTest |      0.758 |      0.280 | 0.663 | 12.536 | 13.745
mod5 | lmerModLmerTest |      0.748 |      0.385 | 0.589 | 12.546 | 13.746
mod6 | lmerModLmerTest |      0.748 |      0.385 | 0.589 | 12.546 | 13.746
mod2 | lmerModLmerTest |      0.739 |      0.185 | 0.680 | 12.530 | 13.745
mod1 | lmerModLmerTest |      0.706 |      0.165 | 0.648 | 13.051 | 13.791
mod3 | lmerModLmerTest |      0.715 |      0.202 | 0.643 | 12.530 | 13.745

Name | AIC weights | AICc weights | BIC weights | Performance-Score
-------------------------------------------------------------------
mod4 |    2.16e-08 |     3.95e-10 |    5.66e-26 |            54.11%
mod5 |    1.43e-06 |     1.93e-08 |    6.44e-25 |            46.74%
mod6 |    1.43e-06 |     1.93e-08 |    6.44e-25 |            46.74%
mod2 |    5.64e-08 |     1.38e-09 |    8.60e-25 |            46.66%
mod1 |       0.986 |        1.000 |        1.00 |            45.63%
mod3 |       0.014 |     4.38e-04 |    1.20e-18 |            36.87%
```



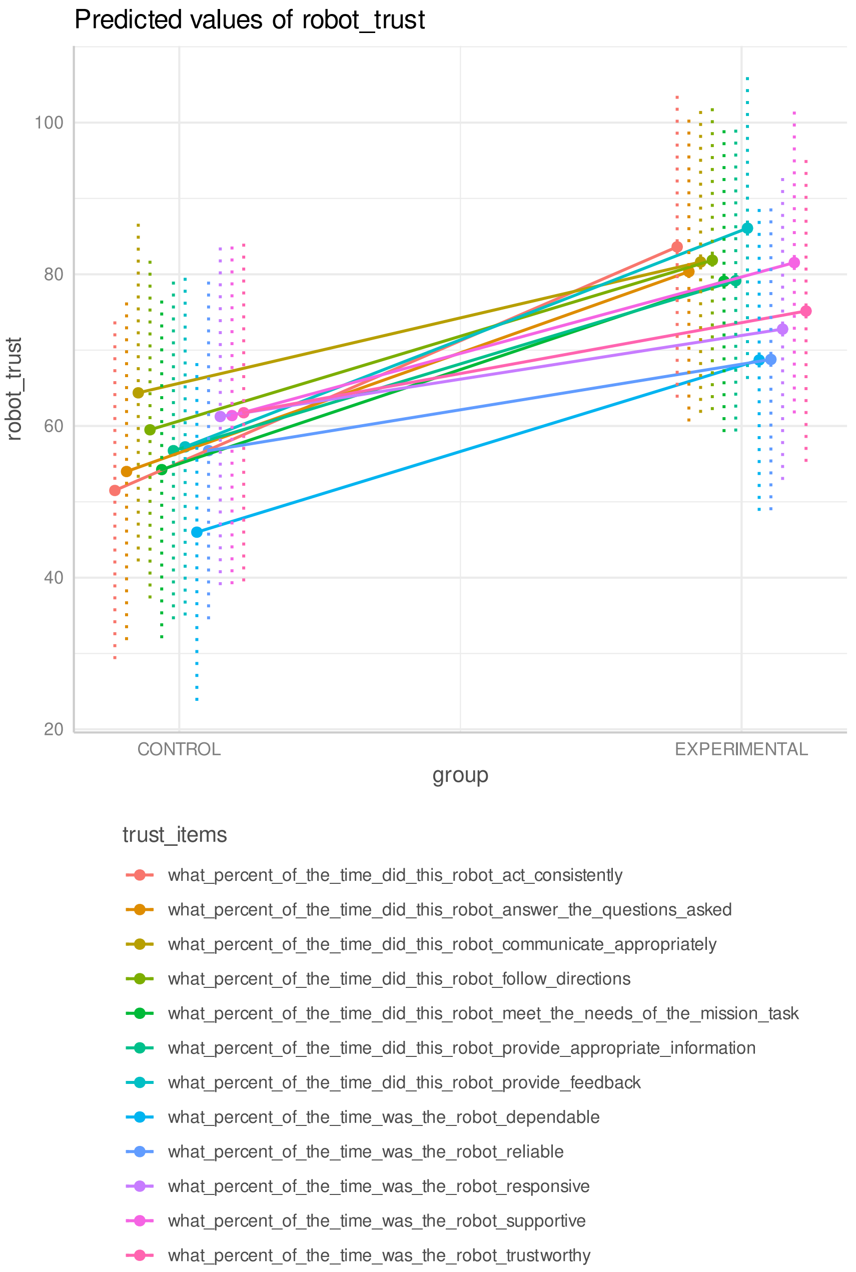
Several Generalized Linear Mixed Effects Models with random intercepts for participant and post-test trust items were fit with fixed effects for treatment group (responsive misty vs control), trust items, and their interaction, controlling for pre-test Negative Attitudes Towards Robots (NARS) and Need For Cognition (NFC) Scores. Even after controlling for language background (native English speaker) and prior experience with robots the treatment effect (responsive versus control) remained significant, predicting higher trust ratings after interacting with the responsive Misty compared to the control Misty (p < 0.002). The best fit is summarized in Table 3. (will add model tests here also, showing improvement of fit with addition of terms and interactions)

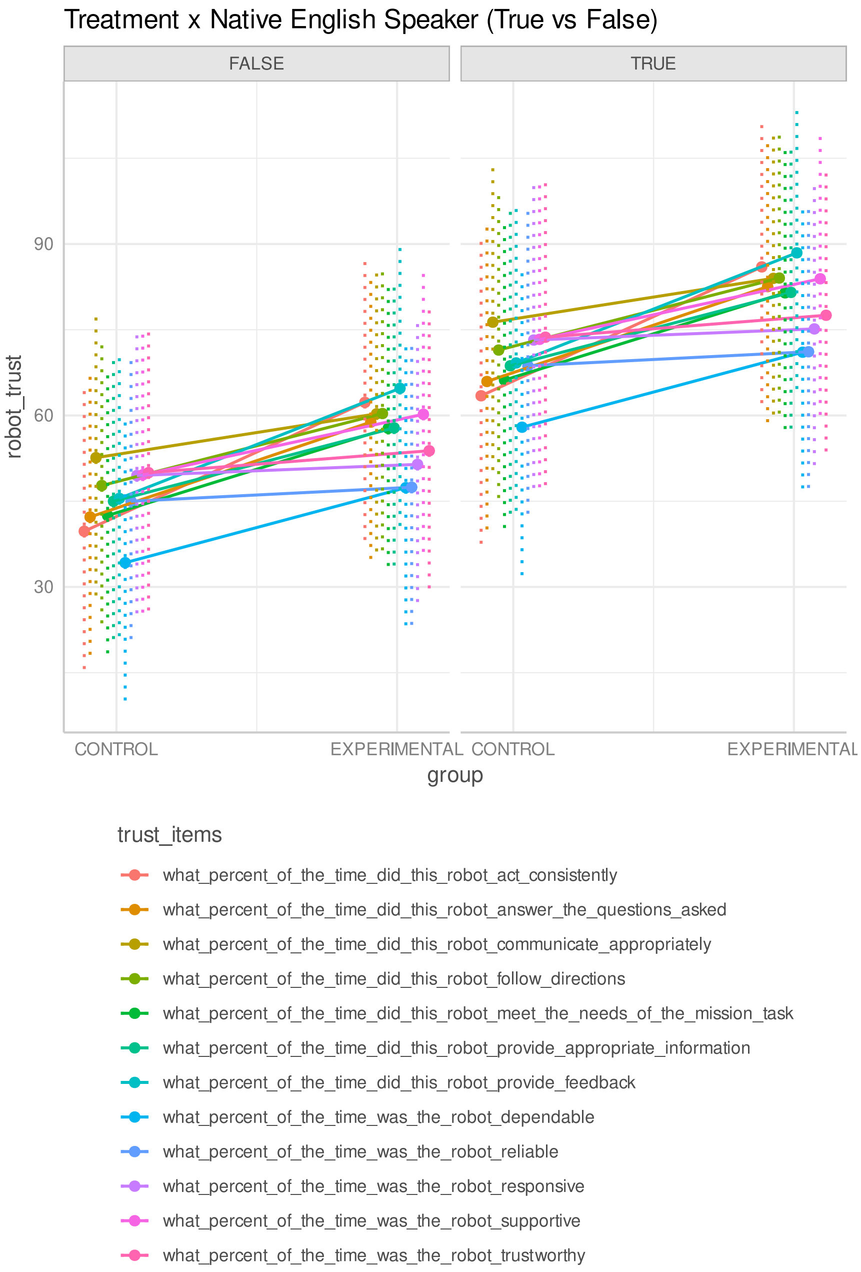
```
robot_trust ~ group * trust_items + nars_pre + nfc_pre + (1 | session_id) + (1 | trust_items)
```

		robot trust	
Predictors	Estimates	CI	p
(Intercept)	65.08	−0.52 – 130.68	0.052
group [EXPERIMENTAL]	26.48	7.21 – 45.76	**0.007**
trust items [what_percent_of_the_time_did_this_robot_answer_the_questions_asked]	2.50	−30.71 – 35.71	0.882
trust items [what_percent_of_the_time_did_this_robot_communicate_appropriately]	12.87	−20.34 – 46.09	0.446
trust items [what_percent_of_the_time_did_this_robot_follow_directions]	8.00	−25.21 – 41.21	0.635
trust items [what_percent_of_the_time_did_this_robot_meet_the_needs_of_the_mission_task]	2.75	−30.46 – 35.96	0.871
trust items [what_percent_of_the_time_did_this_robot_provide_appropriate_information]	5.25	−27.96 – 38.46	0.756
trust items [what_percent_of_the_time_did_this_robot_provide_feedback]	5.75	−27.46 – 38.96	0.733

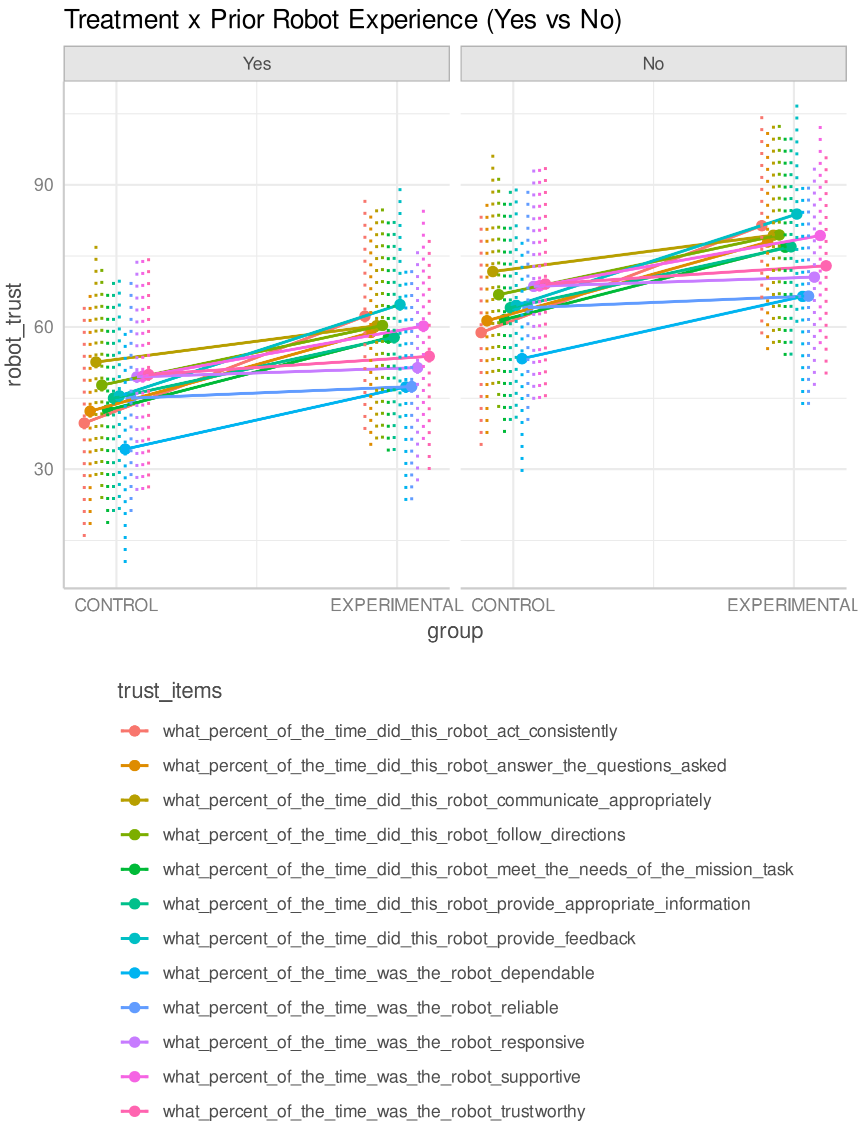
trust items [what_percent_of_the_time_was_the_robot_dependable]	−5.50	−38.71 – 27.71	0.744
trust items [what_percent_of_the_time_was_the_robot_reliable]	5.25	−27.96 – 38.46	0.756
trust items [what_percent_of_the_time_was_the_robot_responsive]	9.75	−23.46 – 42.96	0.563
trust items [what_percent_of_the_time_was_the_robot_supportive]	9.87	−23.34 – 43.09	0.558
trust items [what_percent_of_the_time_was_the_robot_trustworthy]	10.25	−22.96 – 43.46	0.544
native english [TRUE]	21.15	4.71 – 37.59	**0.012**
nars pre	−0.19	−1.15 – 0.77	0.700
nfc pre	−2.56	−11.48 – 6.35	0.572
group [EXPERIMENTAL] × trust items [what_percent_of_the_time_did_this_robot_answer_the_questions_asked]	−5.82	−23.18 – 11.55	0.510
group [EXPERIMENTAL] × trust items [what_percent_of_the_time_did_this_robot_communicate_appropriately]	−14.87	−32.09 – 2.34	0.090
group [EXPERIMENTAL] × trust items [what_percent_of_the_time_did_this_robot_follow_directions]	−9.82	−27.18 – 7.55	0.266
group [EXPERIMENTAL] × trust items [what_percent_of_the_time_did_this_robot_meet_the_needs_of_the_mission_task]	−7.29	−24.50 – 9.93	0.405
group [EXPERIMENTAL] × trust items [what_percent_of_the_time_did_this_robot_provide_appropriate_information]	−9.71	−26.93 – 7.50	0.267
group [EXPERIMENTAL] × trust items [what_percent_of_the_time_did_this_robot_provide_feedback]	−3.29	−20.50 – 13.93	0.707
group [EXPERIMENTAL] ×	−9.42	−26.64 – 7.79	0.282

trust items [what_percent_of_the_time_was_the_robot_dependable]			
group [EXPERIMENTAL] × trust items [what_percent_of_the_time_was_the_robot_reliable]	−20.10	−37.31 – −2.88	**0.022**
group [EXPERIMENTAL] × trust items [what_percent_of_the_time_was_the_robot_responsive]	−20.60	−37.81 – −3.38	**0.019**
group [EXPERIMENTAL] × trust items [what_percent_of_the_time_was_the_robot_supportive]	−11.95	−29.17 – 5.26	0.173
group [EXPERIMENTAL] × trust items [what_percent_of_the_time_was_the_robot_trustworthy]	−18.71	−35.93 – −1.50	**0.033**
Random Effects			
σ^2	188.92		
τ_{00} session_id	254.07		
τ_{00} trust_items	118.38		
ICC	0.66		
N session_id	21		
N trust_items	12		
Observations	250		
Marginal R^2 / Conditional R^2	0.280 / 0.758		


Predicted values of robot_trust
robot_trust
100
80
60
40
20
CONTROL
EXPERIMENTAL
group
trust_items
what_percent_of_the_time_did_this_robot_act_consistently
what_percent_of_the_time_did_this_robot_answer_the_questions_asked
what_percent_of_the_time_did_this_robot_communicate_appropriately
what_percent_of_the_time_did_this_robot_follow_directions
what_percent_of_the_time_did_this_robot_meet_the_needs_of_the_mission_task
what_percent_of_the_time_did_this_robot_provide_appropriate_information
what_percent_of_the_time_did_this_robot_provide_feedback
what_percent_of_the_time_was_the_robot_dependable
what_percent_of_the_time_was_the_robot_reliable
what_percent_of_the_time_was_the_robot_responsive
what_percent_of_the_time_was_the_robot_supportive
what_percent_of_the_time_was_the_robot_trustworthy



Robot experience and native language predictors may overlap, as many of the participants with robot experience are international students.



The greatest predictor of success in completing tasks accurately was not group membership but whether someone was a native english speaker or not, suggesting that language comprehension and the quality of STT translations may have played a larger role in task performance than robot responsiveness. Future work should further explore how language background and robot dialogue systems interact to influence task performance. Solution to this problem could be improved ASR systems or simplified dialogue structures that are less reliant on complex language understanding.

		accuracy	
Predictors	Odds Ratios	CI	p
(Intercept)	230.14	1.82 – 29143.10	**0.028**

group [EXPERIMENTAL]	0.41	0.14 – 1.21	0.107
accuracy items [floor_correct]	1.00	0.22 – 4.45	1.000
accuracy items [status_correct]	0.46	0.11 – 1.93	0.292
accuracy items [suspect_correct]	0.29	0.07 – 1.22	0.092
accuracy items [zone_correct]	0.11	0.02 – 0.52	**0.005**
native english [TRUE]	8.20	1.96 – 34.30	**0.004**
nars pre	0.94	0.87 – 1.00	0.067
nfc pre	0.65	0.34 – 1.25	0.199
Random Effects			
σ^2	3.29		
$\tau_{00\ \text{session\_id}}$	0.10		
$\tau_{00\ \text{accuracy\_items}}$	0.00		
ICC	0.03		
$N_{\text{accuracy\_items}}$	5		
$N_{\text{session\_id}}$	21		
Observations	105		
Marginal R^2 / Conditional R^2	0.331 / 0.351		

```
Random effect variances not available. Returned R2 does not account for random
effects.
Random effect variances not available. Returned R2 does not account for random
effects.
```

```
# Comparison of Model Performance Indices

Name |           Model | R2 (marg.) |  RMSE | Sigma | AIC weights
-----------------------------------------------------------------
mod2 |        glmerMod |      0.338 | 0.425 | 1.000 |       0.993
mod1 |        glmerMod |      0.022 | 0.421 | 1.000 |       0.007
mod5 | lmerModLmerTest |      0.199 | 1.203 | 1.280 |   6.70e-113
mod3 | lmerModLmerTest |      0.187 | 1.207 | 1.287 |   1.30e-113
mod4 | lmerModLmerTest |      0.176 | 1.206 | 1.287 |   7.96e-114

Name | AICc weights | BIC weights | Performance-Score
```

```
-----------------------------------------------------
mod2 |        0.982 |       0.162 |            86.48%
mod1 |        0.018 |       0.838 |            50.42%
mod5 |    7.06e-114 |   3.05e-124 |             9.86%
mod3 |    1.37e-114 |   5.93e-125 |             8.71%
mod4 |    1.15e-114 |   1.83e-124 |             8.17%
```