Analyzing the Fluency of Human-Robot Interactions

Emily Weiss, Zeneve Jacotin, Ryan Blake Jackson, Amy Yuan, and James Boerkoel

Harvey Mudd College
301 Platt Blvd, Claremont, California 91711 USA
{eweiss, zjacotin, bjackson, ayuan, boerkoel}@hmc.edu

Abstract

Fluency—described as the "coordinated meshing of joint activities between members of a well-synchronized team"—is essential to human-robot team success. Human teams achieve fluency through rich, often mostly implicit, communication. A key challenge in bridging the gap between industry and academia is understanding what influences human perception of a fluent team experience to better optimize human-robot fluency in industrial environments. This paper addresses this challenge by developing an online experiment featuring videos that vary the timing of human and robot actions to influence perceived team fluency. Our results support three broad conclusions. First, we did not see differences across most subjective fluency measures. Second, people report interactions as more fluent as teammates stay more active. Third, reducing delays when humans' tasks depend on robots increases perceived team fluency.

Introduction

Human teams achieve fluent interactions in a seemingly effortless fashion through implicit timing signals, such as hesitations or anticipatory actions, or learned conventions, such as turn-taking. These implicit timing cues lead to comfortable, flexible interactions that create a sense of team coherence and fluency. On the other hand, robots are often very precise and rigid in their timing and approach to tasks, with plans optimized for efficiency. This work explores how metrics of the timing of robot actions in a human-robot team correlate with perceived fluency so that we can design scheduling algorithms that better account for humans' natural preferences and social cues for more effective interactions. To our knowledge, these metrics have only been evaluated in the context of human observers of a simulated HRI environment (Hoffman 2019). Our work looks to validate those results when the interactions observed involve actual human and robot agents and extends to two proposed metrics that have not previously been evaluated (Isaacson, Rice, and Boerkoel Jr 2019). We design a set of human-robot interactions using the robot Sawyer[1] that varies key timing metrics. Workers on Amazon's Mechanical Turk (MTurk) platform watch videos of these interactions and respond to statements that aim to capture their perceptions of overall team fluency.

[1]https://www.rethinkrobotics.com/sawyer

Background

The timing of human-robot close collaborative tasks is essential in HRI (Hoffman, Cakmak, and Chao 2014) and has been shown to impact human perceptions of team fluency (Cakmak et al. 2011; Hoffman and Breazeal 2007; Hoffman 2019). Hoffman (2019) introduces human idle time (**HUMAN-IDLE**), robot idle time (**ROBOT-IDLE**), concurrent activity (**CON-ACT**), and functional delay (**FUNC-DELAY**) as quantitative metrics of fluency. Hoffman validated these existing metrics against human perceptions of fluency using a simulated human-robot interaction in which a human and robot completed alternating tasks involving manipulating objects in a shared workspace. Participants watched simulation videos via an online platform and answered questions about how fluent the interactions appeared. Human idle time and functional delay were significantly correlated with the viewers' perception of fluency in an interaction, while robot idle time and concurrent activity were not.

Isaacson, Rice, and Boerkoel Jr (2019) proposed two HRI metrics, *concurrent idleness* and *resource delay*, that they hypothesized will impact human teammates' perceptions of fluency. **Concurrent Idleness (CON-IDLE)** measures the amount of time all agents are simultaneously inactive. **Resource Delay (RESOURCE-DELAY)** is the time difference (positive or negative) between when an agent is ready to use a resource and when that resource becomes available. However, neither of these metrics has been evaluated against humans' perceptions of fluency in human-robot interactions.

Hypotheses

To keep in conversation with prior work, we hypothesize that Hoffman's conclusions about the effect of HUMAN-IDLE (H1), FUNC-DELAY (H2) ROBOT-IDLE (H3), and CON-ACT (H4) will hold (Hoffman 2019) as we transition from a simulated to real human-robot interaction. We also provide hypotheses for the two new metrics. More precisely:

H1. We hypothesize that the sense of fluency will increase as HUMAN-IDLE increases (Hoffman 2019).

H2. We also hypothesize that the sense of fluency will decrease as FUNC-DELAY increases (Hoffman 2019).

Metric	Abbreviation	Subjective Fluency Statements
F1.	[TEAM-FLUENCY]	*The human-robot team worked fluently together.*
F2.	[HUMAN-IMPORTANT]	*The human was the most important member of the team.*
F3.	[ROBOT-TRUST]	*The robot was trustworthy.*
F4.	[ROBOT-UNINTELLIGENT]	*The robot was unintelligent.(R)*
F5.	[ROBOT-UNCOOPERATIVE]	*The robot was uncooperative.(R)*
F6.	[ROBOT-FLUENT]	*The robot contributed to the fluency of the collaboration.*
F7.	[ROBOT-SUCCESS]	*The robot was committed to the success of the team.*
F8.	[ROBOT-IMPORTANT]	*The robot had an important contribution to the success of the team.*

Table 1: Subjective indicators of fluency that appear as part of our participant survey that were taken from Table 1 in (Hoffman 2019). (R) indicates reverse scale (i.e., increased agreement implies decreased perceived fluency) .

H3. We hypothesize that ROBOT-IDLE will not have a significant effect on the perception of fluency (Hoffman 2019).

H4. We hypothesize that CON-ACT will not have a significant effect on the perception of fluency (Hoffman 2019).

H5. We hypothesize that as CON-IDLE increases, the sense of fluency will decrease.

H6. We hypothesize that exchanges will be perceived as more fluent as the absolute value of RESOURCE-DELAY decreases (i.e., resources become ready closer to the time that an agent needs to use them).

Experimental Design

We use Amazon Mechanical Turk (MTurk) and Qualtrics as the platforms for our experiment. One advantage of MTurk is that it reaches a broader demographic sample of the United States population than traditional studies using university students (Crump, McDonnell, and Gureckis 2013). However, it is not entirely free of population biases (Stewart, Chandler, and Paolacci 2017). We restrict participants to MTurk workers at least 18 years of age in the United States with a Human Intelligence Task (HIT) approval rate of at least 99% and at least 50 prior HITs approved. All but one participant completed the survey in less than 30 minutes, and the average amount of time was roughly 12 minutes. We compensate participants $1.50 for submitting the HIT and award a bonus of $1.50 if their submission is usable.

Each participant views two sets of videos in which a human-robot team collaborates to complete a shared workspace task, as shown in Figure 1. We randomize both the sets of videos participants view and the order in which videos appear within each set. Each set contains three videos that display a variation on the metric under observation, with appropriate changes made to the workspace to best test that metric. After viewing each video, the participant answers a series of questions originally from (Hoffman 2019) (see Table 1) asking them to rate their agreement to statements about the fluency of the interaction on 6-point Likert-type items, ranging from strongly disagree (0) to strongly agree (5). Though we recognize the controversy around using parametric statistical methods on this kind of ordinal data, it is widespread to do so, and we undertake our analysis with the knowledge that many studies dating back to the 1930s consistently show that parametric statistics are robust with respect to this practice (Norman 2010).

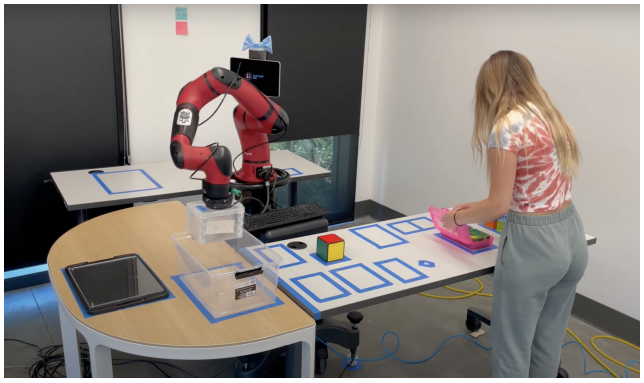

Figure 1: A frame from video E3 (https://youtu.be/IOhh2a5jsuc) demonstrating low resource delay.

Metric	Participants	Video Code & Timing Data		
H1.	49	A1 (86.0%),	A2 (82.8%),	A3 (73.4%)
H2.	53	B1 (7.7%),	B2 (1.7%),	B3 (-5.0%)
H3.	51	C1 (52.7%),	C2 (49.7%),	C3 (44.4%)
H4.	51	D1 (28.2%),	D2 (36.4%),	D3 (48.8)
H5.	51	D1 (5.1%),	D2 (4.3%),	D3 (3.3%)
H6.	52	E1 (24.0%),	E2 (18.0%),	E3 (5.8%)

Table 2: Metric values as a portion of total task time.

We designed five sets of three interactions each using the same shared workspace shown in Figure 1. Each set of interactions attempts to isolate one of the six HRI timing metrics described above. These are reported as a portion of the total interaction time in Table 2). Note that CON-ACT and CON-IDLE are both varied in one set of interactions.

Results

We analyzed our datausing the JASP software package (JASP Team 2022). We use a Bayesian statistical framework for our analysis because 1. the Bayesian approach to statistical analysis provides some robustness to sample size (as it is not grounded in the central limit theorem), 2. the Bayesian approach allows us to examine the evidence both for and against hypotheses (whereas the frequentist approach can only quantify evidence towards rejection of the null hypothesis) (Jarosz and Wiley 2014), 3. the Bayesian approach does not require reliance on p-values used in Null

Hypothesis Significance Testing (NHST) which have come under considerable scrutiny (Berger and Sellke 1987; Simmons, Nelson, and Simonsohn 2011; Sterne and Smith 2001; Wagenmakers 2007), and 4. the rules governing when data collection stops are irrelevant to data interpretation in the Bayesian framework, so it is entirely appropriate to collect data until sufficient evidence has been gathered to draw a meaningful conclusion or until the data collector runs out of time, money, or patience (Edwards, Lindman, and Savage 1963). We use JASP's default general-purpose uninformative prior distributions for all analyses. We follow existing recommendations from in our linguistic interpretations of reported Bayes factors (Bfs) (Jarosz and Wiley 2014).

While the Bayesian statistical approach has become fairly widely used in the Cognitive Science and Psychology research communities, it is still relatively rare in Human-Robot Interaction research. We will, therefore, give a brief, high-level overview to help interpret our quantitative results. The Bfs reported throughout this paper are sometimes referred to as BF_{10}, and are the ratio of the likelihood of the data given the alternative hypothesis to the likelihood of the data given the null hypothesis. BF_{01} shows the opposite ratio, i.e., $BF_{01} = \frac{1}{BF_{10}}$ (Jarosz and Wiley 2014). In general, a Bf < 1 indicates evidence for our null hypothesis (e.g., evidence against a difference in means), a Bf > 1 indicates evidence for our alternative hypothesis (e.g., evidence for a difference in means), and Bf=1 indicates evidence neither for nor against either hypothesis. The further a Bf is from 1, the stronger the evidence and confidence in the corresponding conclusions. A Bf is technically defined as the ratio of probabilities of the data as predicted by each model (e.g., the model that some independent variable had an effect vs. the model that it did not). The decision threshold for deciding whether to accept a model is set by practical considerations, but a Bf > 10 indicates "strong" evidence (Kruschke and Liddell 2018). We analyze responses to the subjective fluency statements delineated in Table 1 via Bayesian repeated measures analyses of variance (RM-ANOVA), and summarize the results for each experiment below.

HUMAN-IDLE Experiment

Participants responded to our eight subjective fluency statements (Table 1) after watching each of three videos across which HUMAN-IDLE varies (videos A1, A2, and A3 in Table 2). According to H1, we would expect the highest perceived fluency associated with video A1 (highest HUMAN-IDLE) and the lowest perceived fluency associated with video A3 (lowest HUMAN-IDLE).

For all subjective fluency statements except F2 [HUMAN-IMPORTANT], a Bayesian RM-ANOVA favors the null model over the model that the video affected the responses, indicating evidence against a difference across videos in perceived fluency. The strength of this evidence ranges from weak (Bf=0.335) to strong (Bf=0.071).

For fluency statement F2, we found substantial evidence supporting an effect of HUMAN-IDLE on participant responses (Bf=6.723). Post hoc tests indicated substantial evidence for a difference between video A1 and A3 (Bf=7.646), with evidence against all other pairwise differences between videos. As shown in Figure 2, agreement was highest when HUMAN-IDLE was lowest. Thus, our data regarding fluency statement F2 are the *opposite* of what H1 predicts.

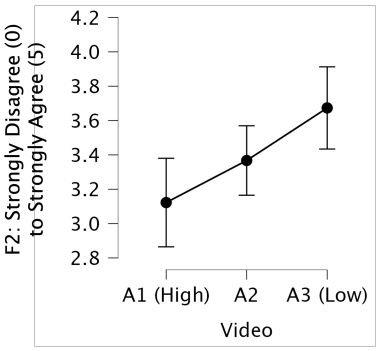

Figure 2: Mean response to fluency statement F2 across videos varying HUMAN-IDLE with 95% credible intervals.

Hoffman's (Hoffman 2019) speculative explanation for H1 was that people might see the human's possibility to rest during idle time as a positive aspect of the collaboration. We do not see variations in agreement with fluency statement F2 as having much bearing on that idea since F2 is a positive statement about the human's contribution rather than a normative one. We also note that F2 was the weakest indicator on the scale in Hoffman's study and was not correlated with any objective metrics, prompting Hoffman to opine that it could reasonably be eliminated from future studies (Hoffman 2019). We have included it here out of interest in perceptions of the human's role. However, it may be less relevant to evaluating our hypotheses than the other seven subjective fluency statements.

FUNC-DELAY Experiment

Participants responded to our eight subjective fluency statements (Table 1) after watching each of three videos varying FUNC-DELAY (videos B1, B2, and B3 in Table 2). According to H2, we would expect the highest perceived fluency after video B3 (lowest FUNC-DELAY) and the lowest perceived fluency after video B1 (highest FUNC-DELAY).

For all subjective fluency statements except F5 and F6, a Bayesian RM-ANOVA favors the null model over the model that the interaction affected the responses, indicating evidence against a difference across interactions in perceived fluency. The strength of this evidence ranges from weak, anecdotal evidence (Bf=0.757) to substantial (Bf=0.101).

For fluency statement F5 [ROBOT-UNCOOPERATIVE], we found very weak evidence supporting an effect of FUNC-DELAY on participant responses (Bf=1.535). Post hoc tests indicated substantial evidence for a difference between video B1 and video B2 (Bf=6.937), with evidence against all other pairwise differences between videos. As shown in Figure 3, agreement with F5 was higher when FUNC-DELAY was more strongly positive (greater *disagreement* with F5 indicates increased perceptions of fluency). We thus conclude that our data regarding fluency statement F5 support H2, at least in the pairwise comparison between the posi-

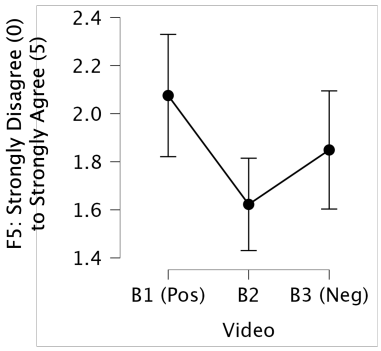

Figure 3: Mean response to fluency statement F5 across videos varying FUNC-DELAY with 95% credible intervals.

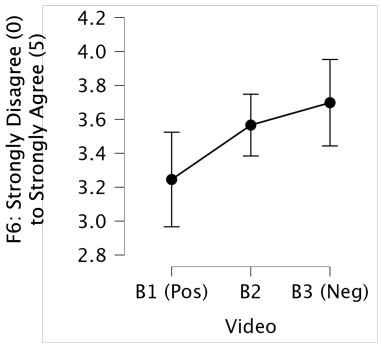

Figure 4: Mean response to fluency statement F6 across videos varying FUNC-DELAY with 95% credible intervals.

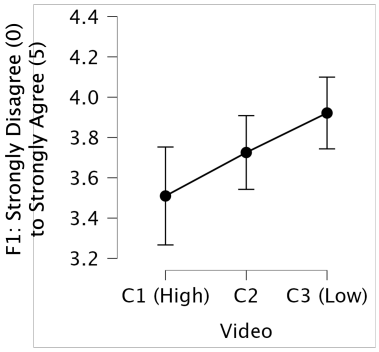

Figure 5: Mean response to fluency statement F1 across videos varying ROBOT-IDLE with 95% credible intervals.

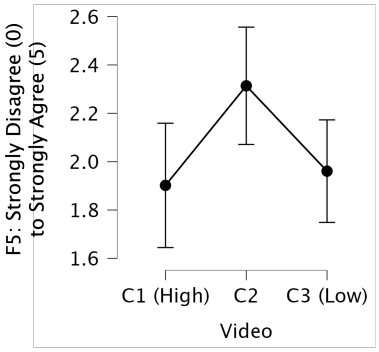

Figure 6: Mean response to fluency statement F5 across videos varying ROBOT-IDLE with 95% credible intervals.

tive FUNC-DELAY condition (video B1) and the near-zero FUNC-DELAY condition (video B2).

We also found very weak evidence supporting an effect of FUNC-DELAY on participant responses to subjective fluency statement F6 [ROBOT-FLUENT] (Bf=1.512). Post hoc tests indicated weak evidence for a difference between video B1 and video B3 (Bf=1.406), with evidence against all other pairwise differences between videos. As shown in Figure 4, agreement with F6 was higher when FUNC-DELAY was more strongly negative. We thus conclude that our data regarding fluency statement F6 weakly support H2.

ROBOT-IDLE Experiment

Participants responded to our eight subjective fluency statements (Table 1) after watching each of three videos across which ROBOT-IDLE varies (videos C1, C2, and C3 in Table 2). According to H3, we expect ROBOT-IDLE not to affect perceived fluency significantly.

Results for most of our subjective fluency statements support H3. Bayesian RM-ANOVAs found weak to substantial evidence against an effect of the videos varying ROBOT-IDLE on F2 (Bf=0.110), F3 (Bf=0.636), F4 (Bf=0.251), F6 (Bf=0.200), F7 (Bf=0.183), and F8 (Bf=0.246). However, there was weak evidence supporting an effect of varying ROBOT-IDLE across videos on fluency statements F1 (Bf=2.090) and F5 (Bf=1.224).

For F1 [TEAM-FLUENCY], post hoc tests revealed substantial evidence (Bf=3.377) for a pairwise difference between only videos C1 and C3 (which have the highest and lowest levels of ROBOT-IDLE respectively). As shown in Figure 5, agreement, and therefore perceived fluency, was highest when ROBOT-IDLE was lowest. Although this contradicts H3, we believe this result intuitively makes sense; the team seemed to work most fluently together when the robot had the least idle time. We emphasize that ROBOT-IDLE was nonzero in all conditions and speculate that perhaps an ROBOT-IDLE of zero would result in a drop in perceived fluency if the robot seemed overworked or unable to complete all of its tasks.

Results showed a different trend for fluency statement F5 [ROBOT-UNCOOPERATIVE]. Post hoc tests showed weak evidence for a difference between videos C1 (high ROBOT-IDLE) and C2 (medium ROBOT-IDLE) (Bf=1.484), and between videos C2 (medium ROBOT-IDLE) and C3 (low ROBOT-IDLE) (Bf=1.703). This result, shown in Figure 6, is more difficult to interpret or explain. However, it is supported by only very weak evidence.

CON-ACT and CON-IDLE Experiment

Participants responded to our eight subjective fluency statements (Table 1) after watching each of three videos across which CON-ACT and CON-IDLE both vary (videos D1, D2, and D3 in Table 2). According to H4, we expect no variation in our subjective fluency statements regardless of the differ-

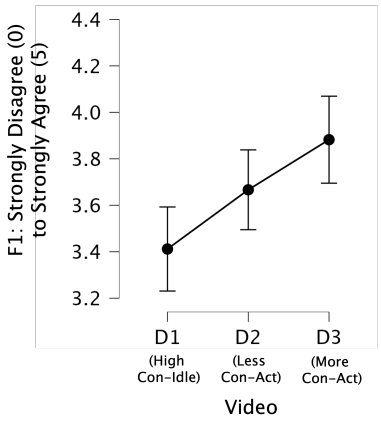

Figure 7: Mean response to fluency statement F1 on videos varying CON-ACT/CON-IDLE with 95% credible intervals.

ences in CON-ACT. However, H5 predicts that the differences in CON-IDLE will lead to increased perceived fluency from video D1 to D2 to D3 as CON-IDLE decreases.

We chose to vary CON-ACT and CON-IDLE both in these videos to maintain realism. We expect these two metrics to be strongly inversely correlated in typical real collaborations. However, this correlation is not necessary, and we can imagine a human-robot team performing a task with 0% of both CON-ACT and CON-IDLE or with 99% CON-ACT and 0% CON-IDLE, for example. Further exploring and disentangling the relationship between CON-ACT and CON-IDLE will require further experimentation in the future. For now, we will attempt to interpret any differences that emerge across our videos with an understanding that we cannot be sure whether they are attributable (more) to changes in CON-ACT or CON-IDLE. However, we are also unsure that this distinction would be meaningful if these metrics tend to be as strongly correlated as we suspect.

For all subjective fluency statements except F1, a Bayesian RM-ANOVA favors the null model over the model that the video affected the responses, indicating evidence against a difference across videos in perceived fluency. The strength of this evidence ranges from weak (Bf=0.526) to strong (Bf=0.076). Thus, H4 is largely supported here.

For fluency statement F1 [TEAM-FLUENCY], a Bayesian RM-ANOVA strongly favors the model that the video affected participant responses (Bf=18.489). Post hoc tests indicated very strong evidence for a difference between video D1 and video D3 (Bf=33.110), with no meaningful evidence for differences between other pairings of the videos. This is the strongest evidence for a difference between conditions found for any variable in this study. As shown in Figure 7, agreement with F1 was higher when CON-ACT was higher and, correspondingly, when CON-IDLE was lower. We interpret this as evidence supporting H5.

RESOURCE-DELAY Experiment

Participants responded to our eight subjective fluency statements (Table 1) after watching each of three videos across

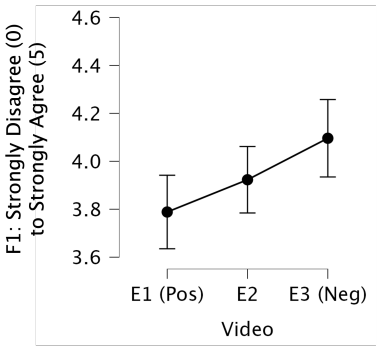

Figure 8: Mean response to fluency statement F1 on videos varying RESOURCE-DELAY with 95% credible intervals.

which RESOURCE-DELAY varies (videos E1, E2, and E3 in Table 2). According to H6, we would expect the highest perceived fluency after video E3 (lowest RESOURCE-DELAY) and the lowest perceived fluency after video E1 (highest RESOURCE-DELAY).

For all subjective fluency statements except F1, a Bayesian RM-ANOVA favors the null model over the model that the video affected the responses, indicating evidence against a difference across videos in perceived fluency. The strength of this evidence ranges from weak (Bf=0.524) to strong (Bf=0.069).

For fluency statement F1 [TEAM-FLUENCY], we found weak evidence supporting an effect of RESOURCE-DELAY on participant responses (Bf=2.101). Post hoc tests indicated substantial evidence for a difference between video E1 and video E3 (Bf=3.700), with evidence against all other pairwise differences between videos. As shown in Figure 8, agreement with F1 was highest when RESOURCE-DELAY was lowest. We thus conclude that our data regarding fluency statement F1 support H6. However, the evidence mentioned above against any differences regarding our other fluency statements concerning H6 gives us pause about concluding overall support for H6; the results are mixed.

Summary of Results

We summarize our findings across our six hypotheses.

H1. [HUMAN-IDLE] Our analysis mostly supports previous findings that human idle time does not significantly impact observer perception of team fluency. The one exception is that we found substantial evidence that the human teammate's contributions were deemed more important as they stayed busier (less idle).

H2. [FUNC-DELAY] Our analysis generally does not support the idea that functional delay significantly impacts an observer's perception of team fluency. The two exceptions are that the robot's cooperativeness and contribution to team fluency *increase* as functional delay decreases. Robots that perform anticipatory actions are seen as more cooperative and more highly contributing to team fluency.

H3. [ROBOT-IDLE] Our analysis supports previous findings that robot idle time does not significantly impact ob-

server perceptions of team fluency in most cases. The one key exception is that we found substantial evidence that observers found the team to be less fluent as the robot was less busy (more idle).

H4. [CON-ACT] & H5. [CON-IDLE] Our analysis generally supports previous findings that concurrent activity does not significantly impact an observer's perception of team fluency. The one notable exception is that team fluency (as measured by subjective fluency statement F1) is perceived to drop significantly as concurrent idleness increases. When combined with our analyses of H1 and H3, a higher-order trend emerges—increases in either teammate's idleness lead to decreased perceived fluency.

H6. [RESOURCE-DELAY] Our analysis found weak support for our hypothesis that reducing resource delay would increase perceived team fluency. However, this result did not persist across all measures of fluency. Combining this with the results from H2 suggests that perceived fluency increases when robots complete tasks in a way that anticipates human actions (e.g., negative functional delay) and complete them in a just-in-time manner (e.g., minimal resource delay).

Conclusions

In this paper, we designed an online experiment that explored how various aspects of timing in human-robot collaboration influence perceived team fluency. Evidence points against differences across most subjective fluency measures. Evidence suggests that perceived team fluency decreases as team members spend more time idle. Finally, the robot is seen as more cooperative and better contributes to team fluency when it completes tasks in an anticipatory, just-in-time manner. This suggests that scheduling human-robot team actions should (1) minimize team member idleness and (2) ensure that transitions, particularly those where human action depends on the robot(s), are well synchronized, e.g., in a way that anticipates human teammates' tasks and avoids creating unnecessary delays. Our results are consistent with prior work based on simulated interactions.

We believe fluency should be a primary consideration for increasing trust and safety as robots become more common in industrial settings. A human worker experiencing their robot teammate acting in a fluent manner can better anticipate the robot's future locations, goals, and tasks, producing long-term trust. This trust increases both the productivity and safety of the human-robot team, which is particularly important within an industrial, manufacturing setting of the type we began to emulate in our experiment. In the future, we aim to conduct an in-person experiment in a more realistic human-robot industrial setting and measure their perceptions of the robot's fluency. Thus, we believe that studying colocated, embodied interactions will allow us to better understand the factors that contribute to perceived team fluency in industrial settings (Xu et al. 2012).

References

Berger, J. O.; and Sellke, T. 1987. Testing a Point Null Hypothesis: The Irreconcilability of p-values and Evidence. *Journal of the American Statistical Association (ASA)*, 82(397).

Cakmak, M.; Srinivasa, S. S.; Lee, M. K.; Kiesler, S.; and Forlizzi, J. 2011. Using spatial and temporal contrast for fluent robot-human hand-overs. In *Proc. of 6th ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI-2011)*, 489–496.

Crump, M. J.; McDonnell, J. V.; and Gureckis, T. M. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PloS one*, 8(3).

Edwards, W.; Lindman, H.; and Savage, L. J. 1963. Bayesian statistical inference for psychological research. *Psychological Review*, 70: 193–242.

Hoffman, G. 2019. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems*, 49(3): 209–218.

Hoffman, G.; and Breazeal, C. 2007. Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. In *Proc. of the 2nd ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI '07)*, 1–8.

Hoffman, G.; Cakmak, M.; and Chao, C. 2014. Timing in human-robot interaction (Workshop). In *Proc. of 9th ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI '14)*, 509–510.

Isaacson, S.; Rice, G.; and Boerkoel Jr, J. C. 2019. MADTN: A tool for measuring fluency in human-robot collaboration. *arXiv preprint arXiv:1909.06675*.

Jarosz, A. F.; and Wiley, J. 2014. What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *The Journal of Problem Solving*, 7.

JASP Team. 2022. JASP (Version 0.16.3)[Computer software].

Kruschke, J. K.; and Liddell, T. M. 2018. Bayesian data analysis for newcomers. *Psychonomic bulletin & review*, 25(1): 155–177.

Norman, G. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15(5): 625–632.

Simmons, J. P.; Nelson, L. D.; and Simonsohn, U. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, (11).

Sterne, J. A.; and Smith, G. D. 2001. Sifting the Evidence – What's Wrong with Significance Tests? *Physical Therapy*, 81(8): 1464–1469.

Stewart, N.; Chandler, J.; and Paolacci, G. 2017. Crowdsourcing Samples in Cognitive Science. *Trends in Cognitive Sciences*.

Wagenmakers, E.-J. 2007. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review*, 14(5): 779–804.

Xu, Q.; Ng, J. S. L.; Cheong, Y. L.; Tan, O. Y.; Wong, J. B.; Tay, B. T. C.; and Park, T. 2012. Effect of Scenario Media on Human-Robot Interaction Evaluation. In *Proc. of the 7th ACM/IEEE Int. Conf. on Human-Robot Interaction (HRI '12)*, 275–276.