A SUPPLEMENTAL MATERIAL

A.1 TOST and BF Equivalence Tests

Usually, scientists try to prove that there is a significant difference or a meaningful effect between two means to show that a new approach is better than an existing one. This is often done with a null hypothesis significance test (NHST), e.g. t-test. In our case, we try to prove that there is no significant difference between participant accuracy and random guessing. Researchers often wrongly conclude that non-significant null hypothesis tests mean that means are the same and that there is no effect [30]. However, the correct way to approach this is to test for equivalence rather than difference.

Statistically, it is impossible to show that an effect δ is exactly 0, meaning there is no difference. But we can define an effect size or equivalence margin $[lb, ub]$ that we consider significant enough to be worth investigating. Any effect δ that falls within the margin can be considered too small to be relevant. Effect sizes can be defined using the standardised Cohen's d [7] or unstandardised raw units (mean difference). For one-sample tests, Cohen's d is given by $d = \frac{|known\ mean - sample\ mean|}{standard\ deviation}$. According to Cohen [7], an effect size of $d = .2$ is considered small, $d = .5$ is considered medium, and $d = .8$ is considered large. However, these definitions are subjective and dependent on the area of research and do not apply in our case.

We consider participant accuracy equivalent to random guessing (50%) as long as it is between 45% and 55%. Thus, our equivalence margin in raw units is $[0.45, 0.55]$. This corresponds to a standardised equivalence margin of $[-0.42, 0.5]$ using Cohen's d. Two ways to test for equivalence are e.g., the frequentist two one-sided t-tests (TOST) procedure and the Bayes factor interval null procedure [30, 32]. The TOST procedure has the following null and alternative hypotheses: $H_0 : \delta \leq lb\ OR\ \delta \geq ub$ and $H_1 : \delta > lb\ AND\ \delta < ub$.

The Bayes factor (BF) interval null procedure is based on Bayes' rule and compares two hypotheses using the Bayes factor. The BF_{10} indicates how much more likely the data are under H_1 than H_0 [32] and gives a magnitude of the evidence for equivalence [26]: $1 < BF < 3$ (anecdotal evidence), $3 < BF < 10$ (moderate evidence), $10 < BF < 30$ (strong evidence), $30 < BF < 100$ (very strong evidence), $BF > 100$ (extreme evidence). The posterior largely depends on the selection of the prior. Opinions are divided between using subjective or objective priors [32]. We used an objective prior since we do not have prior information about the effect size of our data, to make the results more reproducible, and because we want to minimize the influence of the prior on the posterior [38]. A commonly used prior is the Cauchy distribution centered at 0. It is similar to the Normal distribution making small effect sizes more likely than larger effect sizes. Unlike the Normal distribution, the Cauchy distribution has fatter tails, making large effect sizes still possible albeit less plausible [38, 53].

Linde et al.[32] recommend the Bayes factor procedure for small sample sizes instead of the TOST procedure as TOST is maximally conservative and does not discriminate well for small equivalence margins. Nevertheless, we performed the TOST procedure to provide a p-value in our analysis, as our equivalence margins are not small ($d > .2$). We also computed the Bayes factor which offers more evidence for equivalence than the p-value [16].

A.2 Additional Results Preference Round

A.2.1 Actor 1 vs. Actor 2. Paired-samples t-test: In the VR study, there were two outliers in the difference scores which resulted in a non-normal distribution of the data (Shapiro-Wilk's test with outliers: $p = .021$, Shapiro-Wilk's test without outliers: $p = .598$). Without outliers, Actor 2 (59.19% ± 13.32%) was preferred significantly more with 18.39% than Actor 1 (40.81% ± 13.32%), $t(39) = 4.367, p < .001, d = .69$. With outliers in the data, Actor 2 (65.77% ± 17.09%) was preferred significantly more with 13.54% than Actor 1 (43.23% ± 17.09%), $t(41) = 2.567, p = .014, d = .396$.

A.2.2 1st vs. 2nd Recording in Set. We theorized that participants might tend to choose the second recording of the set more often because it might be fresher in their minds than the first video. However, for the Web study, the paired t-test was not significant, $t(46) = .896, p = .375, d = .131$. The data in the VR study were not normally distributed ($p = .034$). Therefore, we performed a Wilcoxon signed-rank test in addition to the paired t-test. Participants selected the second recording in 6.5 ± 1.55 sets, and the first

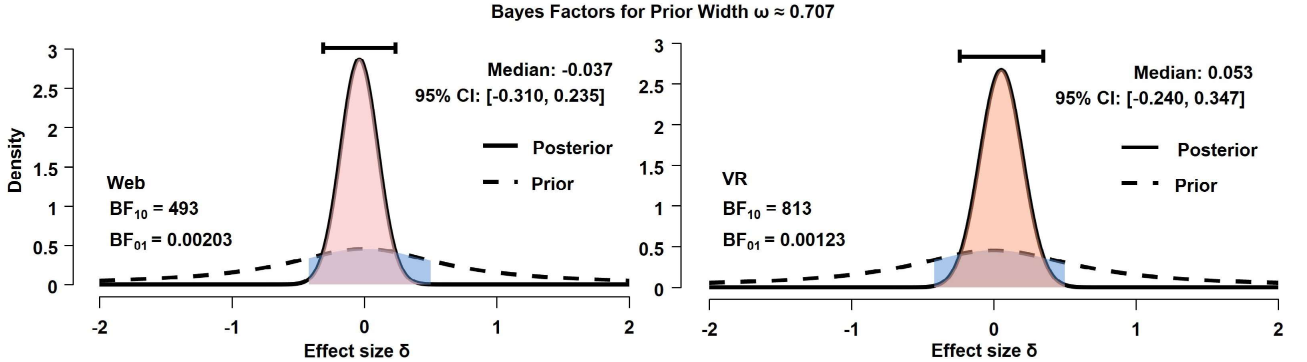

Figure A.1: Bayes factors for Web and VR study for prior width $\omega = 0.707$ and posterior distributions with median and 95% credible interval. The prior probability under H_1 (equivalence) is shown in blue, and the posterior probability under H_1 is shown in red. The remaining uncolored area of the prior is the prior probability for H_0 (no equivalence).

recording in 5.5 ± 1.55 sets. This was a small but significant difference, $t(41) = 2.091, p = .043, d = .323$. Of 42 participants, 19 participants selected the second recording more often, 12 participants the first recording, and 11 participants selected both recordings equally often. The Wilcoxon signed-rank test was significant, but there was no difference in medians between the number of first recordings selected and the number of second recordings selected, $z = 2.08, p = .038$.

Table A.1: Recording procedure for creating a consistent dialogue dataset from single- and multi-user recordings. The position of the first speaker (L = left, R = right) alternates to make sure that the actors do not stand in the same position all the time. The first speaker is always the one who records the dialogue.

Dialogues	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Position of First Speaker	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R
First Speaker (Round 1)	Actor 1										Actor 2									
Single-User	Actor 1/Actor 2																			
First Speaker (Round 2)	Actor 2										Actor 1									

Table A.2: Order of recordings (Rec) for Preference and Detection Round. We used a Balanced Latin square design to create a pseudo-random order of single(Sgl)- and multi(Mlt)- user recordings. In the Preference Round, the dialogues (Dlg) in each set are the same. In the Detection Round, the order of the dialogues has been chosen such that there are at least 2 different dialogues between recurring dialogues. The dialogues and corresponding recordings are highlighted in different colors.

(a) Preference Round

Sets	Mlt/ Sgl	Rec	Dlg	Sets	Mlt/ Sgl	Rec	Dlg
Set 1	M	R1	1	Set 7	S	A2	2
	S	A2	1		M	R1	2
Set 2	M	R2	2	Set 8	M	R1	3
	S	A1	2		S	A1	3
Set 3	S	A2	3	Set 9	S	A2	4
	M	R2	3		M	R2	4
Set 4	S	A1	4	Set 10	S	A1	5
	M	R1	4		M	R1	5
Set 5	M	R2	5	Set 11	M	R1	6
	S	A2	5		S	A2	6
Set 6	S	A1	6	Set 12	M	R1	1
	M	R2	6		S	A1	1

(b) Detection Round

#	Mlt/ Sgl	Rec	Dlg	#	Mlt/ Sgl	Rec	Dlg
1	S	A1	1	13	S	A1	5
2	M	R1	2	14	M	R2	1
3	M	R1	3	15	S	A2	3
4	S	A2	4	16	M	R1	4
5	M	R2	5	17	M	R2	2
6	S	A1	2	18	S	A1	6
7	S	A2	6	19	S	A2	1
8	M	R2	3	20	M	R2	4
9	M	R1	5	21	S	A1	3
10	S	A1	4	22	M	R1	1
11	M	R2	6	23	M	R1	6
12	S	A2	2	24	S	A2	5

Table A.3: List of deciding factors named by participants for Preference and Detection Rounds during the Web and VR study. Similar and more prominent factors that were mentioned in both studies are highlighted in different colors.

Web Preference Round		Web Detection Round		VR Preference Round		VR Detection Round	
Deciding Factors	**# Participants**	**Deciding Factors**	**# Participants**	**Deciding Factors**	**# Participants**	**Deciding Factors**	**# Participants**
Body language, Gestures, Interactions	21	Tone, Pitch, Volume, Flow of conversation	38	Hand Movements, Animation	22	Sound, Tone, Pitch, Volume of voice	35
Naturalness, Human-like voices	20	Accent, Pronunciation, Intonation	13	Naturalness	17	Accent, Pronunciation	15
Clarity of voice, Quality of speech	18	(Hand) Movements	6	Clarity of voice, Pronunciation	14	Reactions, (Hand) Movements	11
Speaking speed	13	Voice editing/effects, distortions	4	Tone and Volume	13	Speaking speed	6
Tone, Pitch, Volume	10	Breathing	3	Energy and emotions, Interesting conversations	13	Voice editing/effects	4
Smoothness, Fluency, Latency, Avatars interrupting	8	Speaking speed	3	Speaking speed	12	Response speed, Interruptions	4
Gestures/Voice matching topic of conversation	7	Avatar appearance	1	Gesture/Voice matching topic of conversation	8	Naturalness, Robot-like	3
Expressivity, Emotions	4	Energy when speaking	1	Distance between avatars	5	Clarity of voice	1
Distance between avatars	1	Distance between avatars	1	Intonation	4		

A.3 Additional Results Detection Round

A.3.1 TOST Equivalence Test. In the VR study, the participant accuracies were not normally distributed according to Shapiro-Wilk's test ($p = .033$). However, the inspection of a Q-Q plot indicated near normality and there were no outliers in the data which is why we also performed a parametric TOST: Participants' detection accuracies were not significantly higher by .006 (95% CI, −.031 to .043) than the guessing average 0.5, $t(41) = .324,\ p = .748$, Cohen's $d = .05$ (95% CI −.253 to .352). Participants' accuracies were significantly higher by .056 than the lower bound .45, $t(41) = 3.044,\ p = .002$, Cohen's $d = .47$ (95% CI .141 to .786), and significantly lower by .044 than the upper bound .55, $t(41) = -2.396,\ p = .011$, Cohen's $d = .37$ (95% CI .055 to .68). The resulting 90% CI (because of $2\alpha = 0.1$) is $[.475, .537]$ which lies within our equivalence bounds $[0.45, 0.55]$.

A.3.2 Effect of Distance on Detection Accuracies and Perception of Politeness. We gathered additional data during the VR study only. Participants were watching the recordings from 3 different distances (between-subjects condition) based on proxemics (personal, social, and public distance), and we wanted to know whether distance would affect participants' detection accuracy and their perception of politeness of the avatars. We conducted a one-way ANOVA. Participants were divided into three different conditions: Personal ($n = 14$), Social ($n = 14$), and Public ($n = 14$). The results were normally distributed (Shapiro-Wilk's tests: $p = .758,\ p = .119,\ p = .054$, Personal, Social, Public respectively) and there was homogeneity of variances (Levene's test: $p = .931$). Detection accuracies slightly decreased from the Personal condition(.515 ± .130, mean ± standard deviation) to the Social condition (.503 ± .115) and stayed the same for the Public condition (.503 ± .119). There was no statistically significant difference in accuracies, $F(2, 39) = .045,\ p = .956$. We performed a Kruskal-Wallis H test to determine if there were differences in politeness/ignoring scores for different distance conditions ($n = 14$ in all distance conditions). Distributions of politeness and ignoring scores were not similar for all conditions, therefore, we reported the mean ranks for each condition instead of the median. Mean ranks of the politeness scores were very similar for each condition (Personal: 21.68, Social: 20.93, Public: 21.89). Ignoring scores were similar for the Personal and Social conditions (22.25 and 25.96 respectively) and slightly lower for the Public condition (16.29). The mean ranks were not statistically significantly different between the three conditions, $\chi^2(2) = .049,\ p = .976$ (politeness), and $\chi^2(2) = 4.722,\ p = .094$ (ignoring).