Now Look Here! ⇓ Mixed Reality Improves Robot Communication Without Cognitive Overload

Nhan Tran[1,2], Trevor Grant[3], Thao Phung[2], Leanne Hirshfield[3], Christopher Wickens[4], and Tom Williams[2]

[1] Cornell University, Ithaca NY 14853, USA nt322@cornell.edu Colorado School of Mines, Golden CO 80401, USA twilliams@mines.edu
[2] University of Colorado Boulder, Boulder CO 80309
leanne.hirshfield@colorado.edu
[3] Colorado State University, Fort Collins CO 80523

Abstract. Recently, researchers have initiated a new wave of convergent research in which Mixed Reality visualizations enable new modalities of human-robot communication, including *Mixed Reality Deictic Gestures* (MRDGs) – the use of visualizations like virtual arms or arrows to serve the same purpose as traditional physical deictic gestures. But while researchers have demonstrated a variety of benefits to these gestures, it is unclear whether the success of these gestures depends on a user's level and type of cognitive load. We explore this question through an experiment grounded in rich theories of cognitive resources, attention, and multi-tasking, with significant inspiration drawn from *Multiple Resource Theory*. Our results suggest that MRDGs provide task-oriented benefits regardless of cognitive load, but only when paired with complex language. These results suggest that designers can pair rich referring expressions with MRDGs without fear of cognitively overloading their users.

Keywords: Mixed Reality · Cognitive Load · Deictic Gesture

1 Introduction

Successful human-robot interaction in many domains relies on successful communication. Accordingly, there has been a wealth of research on enabling human-robot communication through natural language [31, 49]. However, just like human-human dialogue, human-robot dialogue is inherently multi-modal, and requires communication channels beyond speech. Human interlocutors regularly use gaze and gesture cues to augment, modify, or replace their natural language utterances, and will often use deictic gestures such as pointing, for example, to (1) direct interlocutors' attention to objects in the environment, (2) reduce the number of words that the speaker must use to refer to their target referents, and (3) lower the cognitive burden imposed on listeners to interpret those utterances.

Due to the prevalence and utility of deictic gestures in situated communication, human-robot interaction researchers have sought to enable robots to understand [30] and generate [42, 40, 39] deictic gestures as humans do. However, the ability to understand and generate deictic gestures comes with hardware requirements that can be onerous or unsatisfiable in certain use cases. While perceiving deictic gestures only requires a camera or depth sensor, generating deictic gestures requires a specific morphology (e.g., expressive robotic arms). This fundamentally limits gestural capabilities, and thus overall communicative capabilities, for *most* robotic platforms in use today. As examples, robots such as mobile bases used in warehouses, assistive wheelchairs, and unmanned aerial vehicles (UAVs) lack the morphologies needed to effectively communicate in this manner. Even for robots that do have arms, traditional deictic gestures have fundamental limitations. In contexts such as urban or alpine search and rescue, for example, robots may need to communicate about hard-to-describe and/or highly ambiguous referents in novel, uncertain, and unknown environments.

Consider, for example, an aerial robot in a search and rescue context. If the robot needs to generate an utterance such as "I found a victim behind *that tree*" (cf. [68]), the ability to precisely pick out the target tree using a gestural cue would be of great value, as the referring expressions the robot would need to generate without using gesture would likely be convoluted (e.g., "the fourth tree from the left in the clump of trees to the right of the large boulder") or not readily human-understandable (e.g., "the tree 48.2 meters to the northwest").

Unfortunately, such a UAV would be unlikely to have an arm mounted on it solely for gesturing, meaning that physical gesture is not a realistic possibility, no matter its utility. Moreover, even in the unlikely case that the robot had an arm mounted on it, it is unlikely that a traditional pointing gesture generated by such an arm would be able to pinpoint a specific far-off tree. In this work, we present a solution to this problem that builds on recent collaborative work between the HCI subfields of Mixed Reality and Human-Robot Interaction, which have come together to initiate a new wave of convergent research in which Mixed Reality visualizations are used to enable fundamentally new modalities of human-robot communication. Specifically, we present a *Mixed Reality* solution that enables robots to generate effective deictic gestures without imposing any morphological requirements. Specifically, we present the first use of the Mixed Reality Deictic Gestures *MRDGs* proposed by Williams et al. [67] to be deployed in a rich, multimodal, task-based environment using real robotic and mixed reality hardware.

MRDGs are visualizations that can serve the same purpose as traditional deictic gestures, and which fall within the broad category of *view-augmenting* mixed reality interaction design elements in the Reality-Virtuality Interaction Cube framework [65]. Williams et al. [67] divide these new forms of visual gestures into *perspective-free* gestures that can be projected onto the environment, and *allocentric* gestures (visualized in the perspective of the listener) that can be displayed in teammates' augmented reality (AR) head-mounted displays. Recent work on perspective-free gestures has focused on the *legibility* of projected gestures [54], while recent work on allocentric gestures has focused on gesture effectiveness when paired with different kinds of language (in virtual online testbeds) [63, 64] and on effectiveness of *ego-sensitive allocentric* gestures such as virtual arms [18, 7, 20, 16, 14]. In this work we focus on this first, (non-egosensitive) allocentric category of MRDG.

In previous work, Williams et al. [64] (see also [63]), suggested that (non-ego-sensitive) allocentric MRDGs might increase communication accuracy and efficiency, and, when paired with complex referring expressions, might be viewed as more effective and likable. However, MRDGs have been primarily tested in video-based simulations [64, 63], or in rigid experiments with low ecological validity [18, 7, 20]. In this paper, we present the first demonstration of MRDGs generated on actual AR Head-Mounted Displays (the Hololens) by commercial-grade robots, in rich, multi-modal, task-based environments.

Deploying MRDGs in these realistic task-based robotic environments allows us to how the dimensions of realistic task contexts and realistic robotic communication may or may not actually afford the effective use of such gestures. As previously pointed out by Hirshfield et al. [23], the tradeoffs between language and visual gesture may be highly sensitive to teammates' level and type of cognitive load. For example, Hirshfield et al. [23] suggest that it may not be advantageous to rely heavily on visual communication in contexts with high visual load, or to rely heavily on linguistic communication in contexts with high auditory or working memory load. These intuitions are motivated by prior theoretical work on human information processing, including the Multiple Resource Theory (MRT) by Wickens [57, 58]. On the other hand, recent work conducted in rigid, non-task-based laboratory studies involving robots with *purely* gestural capabilities has demonstrated the extremely successful effectiveness of MRDGs at manipulating interactant attention in order to maximize object task-based metrics of interaction success [18, 7].

It is thus unclear whether the success of MRDGs depends on the level and type of cognitive load that a user is under, or the type of multimodal communications strategies they are used in service of, or whether they might simply be broadly effective regardless of these factors. In this work, we thus analyze the use of MRDGs in the context of different multimodal robot communication strategies through a human-subjects experiment whose experimental design is grounded in rich theories of cognitive resources, attention, and multi-tasking, with significant inspiration drawn from *Multiple Resource Theory.*

Our results provide partial support for a *Universal Benefit Hypothesis*, which suggests that MRDGs provide task-oriented benefits regardless of what type of load users are under; our results show that MRDGs may only provide these benefits when paired with rich referring expressions. These results provide critical insights for designers, suggesting that designers operating in Mixed Reality Robotic domains can and should pair rich referring expressions with MRDGs without fear of cognitively overloading their users in certain cognitive contexts.

The rest of the paper proceeds as follows. In Section 2, we discuss related work on Mixed Reality HRI and the resource theories of attention and multitasking. In Section 3, we present a human-subject experiment to study the effectiveness of different robot communication styles under different types of cognitive load. In Section 4, we present the results of this experiment. Our results show that MRDGs enhance the effectiveness of robot communication, regardless of how robots' verbal communication is phrased, and regardless of what level and type of mental workload interactants are under (at least under the phrasings and parameterizations used in this experiment). Finally, in Sections 5 and 6 we conclude with general discussion, and recommendations for future research.

2 Related Work

2.1 AR for HRI

Mixed reality technologies that integrate virtual objects into the physical world have sparked recent interest in the Human-Robot Interaction (HRI) community [66] because they enable better exchange of information between people and robots, thereby improving mental models and situation awareness [47].

Despite significant research on augmented and mixed reality for several decades, [4, 3, 50, 69, 5] and acknowledgement of the potential for impact of AR on HRI [15, 32], only recently has there been significant and sustained interest in the Virtual, Augmented, and Mixed Reality for Human-Robot Interaction (VAM-HRI) community [66, 17, 53]. Recent works in this area include approaches using AR for robot design [36], calibration [43], and training [46]. Moreover, there are a number of approaches towards communicating robots' perspectives [22], intentions [2, 13, 8, 11, 9], and trajectories [6, 52, 37, 12].

Sharing perspectives is one of the best ways to improve human-robot interaction. Amor et al. [1] suggest that projecting human instructions and robot intentions in a constrained and highly structured task environment improves human robot teamwork and produces better task results [1, 2, 13]. Similarly, Sibirtseva et al. [44], enable robots receiving natural language instructions to reflexively generate mixed reality annotations surrounding candidate referents as they are disambiguated [44]. Finally, several researchers [63, 64, 19, 7, 20, 14] investigate AR augmentations as an *active* rather than passive communication strategy, generated as gestures accompanying verbal communication.

2.2 Resource Theories of Attention and Multitasking

The previous section outlines the current state of AR for HRI, especially with respect to active and passive communication. We argue that future robots must tailor visual cues to the contextual needs of human teammates. As the first step towards enabling adaptive multimodal interfaces for human-robot communication, this study aims to unravel the interaction between MRDGs, human mental workload, and the nature of the multimodal interface in which gestures are generated, to determine whether mental workload and multimodality should be accounted for in future adaptive systems. The theoretical foundation for our investigation is supported by theories of attention and multitasking, especially as they pertain to mental workload and multiple resources [58–60].

First, resource theory posits limits to multitasking related to the difficulty or mental workload imposed by a task, and the relation between the resources demanded by the task (MWL) and the cognitive resources available to the user [61]. In a dual task context, when one (primary) task is increased in difficulty, the resources available for a secondary task decrease, along with performance on that task, in a *reciprocal* fashion. This is the foundation of *single resource theory* [25, 35]. In our experiment, we assess the MWL demands of a primary robotics task by a standardized scale, the NASA Task Load Index (TLX).

Second, the theory of resources in multi-task contexts has been expanded to assume *multiple resources* [34] defined on the basis of neurophysiological structures such as the auditory versus visual cortex or the spatial and verbal cerebral hemispheres [62, 59, 60]. As applied to multitasking, the existence of multiple resources implies that the perfect reciprocity between the demands of one (primary) task, and a concurrently performed (secondary) task no longer holds, to the extent that the two tasks employ different resources (e.g., auditory presentation on one, visual on the other). Performance on one task can still be preserved, despite higher demands on the other. This high time-sharing efficiency when separate resources are used, forms the basis of our empirical work, and our envisioned future adaptive interfaces: to switch the modality of information provided in a dialogue, as a function of the higher demands of a primary task. In the next two subsections, we provide further detail on relevant prior work on both Multiple Resource Theory and on Theories of Dual-Tasking.

2.3 Multiple Resource Theory

The Multiple Resource Theory (MRT) proposed by Wickens [57, 58] states that people have different cognitive resources for processing information. These resources can process different information at the same time and can be categorized along three dimensions: 1) early vs. late processing stage, 2) spatial vs. verbal processing code, and 3) visual vs. auditory modality [58].

The complexity of the tasks determines how these resources are utilized. For example, if the various tasks need to tap the same pool of resource, it will process the information sequentially. If the tasks need to access different resources, information will be processed in parallel. Additionally, the task performance indicates how these resource limits are reached. When two or more tasks that require a single resource are performed at the same time, a supply and demand occurs. Task error and performance decrement occur when a task that requires the same resource causes excess workload. Furthermore, MacDonald et al. [28] suggests that there is a complex relationship between workload and job performance: An increase in workload does not always result in a decrease in performance and performance can be affected by both high and low workload [28]. If the users are under low workload, also known as underload, they might become bored, lose situation awareness, and reduce alertness.

Applying MRT in the context of collocated human-robot teaming, it is even more crucial for robots to communicate using the appropriate modalities and context-aware methods that do not overload the mental resources of the human operator. Wickens' MRT framework can be used to evaluate: (1) when tasks can be carried out simultaneously, (2) how tasks interfere with each other, and

(3) how increasing in one task's difficulty may impact other task's performance [58]. Our current work is motivated by a vision of future MRT-inspired robotic communication systems that could be used in scenarios involving multitasking by human operators and multimodal presentation of information. Such adaptive systems can potentially lead to more efficient use of resources, more task-relevant presentation of information, increased task performance, improved perception of the robot, and safer environment for collocated human-robot collaboration.

2.4 Dual-Tasking

Many researchers have used dual-tasking to study the limitations of human ability to process information [45, 41]. In the dual-task method, subjects perform two tasks concurrently; often one of thee is designated primary and the other secondary. It is assumed (and instructed) that the participant will allocate necessary resources to the primary task, so that it does not deteriorate in the presence of the secondary task. The results then provide insights into how tasks can be carried out together and how they contribute to the workload. A classic example is driving a car and talking with passengers. This dual-task situation becomes challenging when the demand for driving task increases in poor road conditions, or if the secondary task involves a heated argument [45].

Wickens [58] demonstrated how performance decreases under dual-task condition and provides theoretical implications for resource allocation. For example, two tasks requiring the same modality will produce lower performance compared to when modalities differ. During an intense car drive that requires the driver to increase demand in processing the primary task spatially, it would be much easier to verbally process a secondary task (e.g., additional navigation instructions). Taking into account the impact of dual-task performance on mental workload, we designed a dual-task experiment that systematically varies cognitive load by changing the input modality (visual vs. auditory presentation of task structure) and the central processing code (spatial annotation over the target object vs. verbal instruction describing where the target object is located).

3 Experiment

We experimentally assessed whether the level and type of cognitive load and/or the multimodal communication strategies into which MRDGs are integrated mediate the effectiveness of those mixed-reality deictic gestures. To do so, our experiment used a 4×3 within-subjects experimental design (as described below), in which four levels and types of of cognitive load (high visual perceptual load, high auditory perceptual load, high working memory load, and low overall load) were crossed with three different multi-modal communication strategies (MRDGs paired with complex vs simple language, as well as vs complex language alone without the use of mixed-reality deictic gestures). This design is based on the assumptions that there are different perceptual resources, that MRDGs employ visual-spatial resources in accordance with MRT, and that the

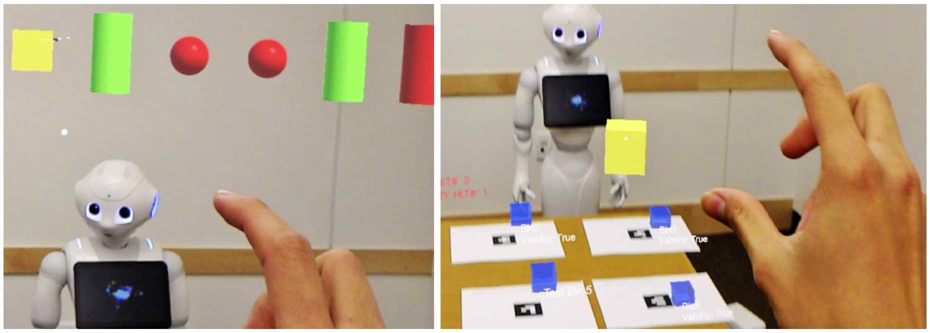

Fig. 1. During the experiment, participants play a mixed reality game using the Microsoft HoloLens. The Pepper robot is positioned behind the table, ready to interact.

linguistic dimensions of different communication strategies differentially employ auditory resources in accordance with MRT. Our experiment was designed to contrast two overarching competing hypotheses.

The first hypothesis, the *Cognitive Contextual Benefit Hypothesis*, formalizes the intuitions of Hirshfield et al. [23]:

H1.1 Users under high **visual perceptual load** will perform quickest and most accurately when robots rely on complex language without the use of MRDGs.
H1.2 Users under high **auditory perceptual load** will perform quickest and most accurately when robots rely on MRDGs without the use of complex language.
H1.3 Users under high **working memory load** will perform quickest and most accurately when robots rely on MRDGs without the use of complex language.
H1.4 Users under **low overall load** will perform quickest and most accurately when robots rely on MRDGs paired with complex language.

The second hypothesis, the *Universal Benefit Hypothesis*, instead would suggest that due to the substantial task-oriented benefits provided by mixed reality decitic gestures (as observed in experimental work on real robotic and MR hardware published after that Hirshfield et al. [23], e.g. [18]), MRDGs will be universally beneficial, regardless of cognitive load.

H2 Mixed-reality deictic gestures will be equally effective regardless of level and type of cognitive load.

3.1 Task Design

We will now describe the design of the experimental task designed to assess these two competing hypotheses. Participants interacted with a language-capable robot while wearing the Microsoft HoloLens over a series of trials, with the robot's communication style and the user's cognitive load systematically varying

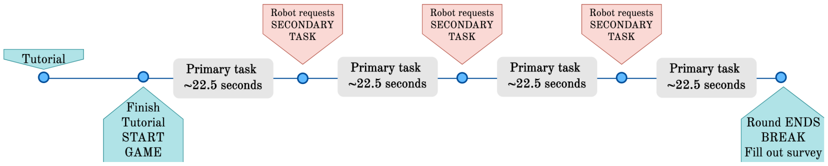
Tutorial
Finish Tutorial START GAME
Primary task ~22.5 seconds
Robot requests SECONDARY TASK
Primary task ~22.5 seconds
Robot requests SECONDARY TASK
Primary task ~22.5 seconds
Robot requests SECONDARY TASK
Primary task ~22.5 seconds
Round ENDS BREAK Fill out survey

Fig. 2. After completing the tutorial and familiarizing themselves with the HoloLens, participants engage in each of the twelve trials. Their primary task is to pick-and-place the target block into the target bin. Throughout a 90-second experiment trial, the robot Pepper interrupts every 22.5 seconds with a secondary task.

between trials. The experimental task ensemble employed a dual-task paradigm oriented around a tabletop pick-and-place task. Participants view the primary task through the Microsoft HoloLens, allowing them to see virtual bins overlaid over the mixed reality fiducial markers on the table, as well as a panel of blocks above the table that changes every few seconds. As shown in Fig. 1, the Pepper robot is positioned behind the table, ready to interact with the participant.

Primary Task

The user's *primary task* is to look out for a particular block in the block panel (selected from among *red cube*, *red sphere*, *red cylinder*, *yellow cube*, *yellow sphere*, *yellow cylinder*, *green cube*, *green sphere*, *green cylinder*[4]). These nine blocks were formed by combining three colors (red, yellow, green) with three shapes (cube, sphere, cylinder). Whenever participants see this target block, their task is to pick-and-place it into any one of a particular set of bins. For example, as the game starts, the robot might tell a user that whenever they see a *red cube* they should place it in bins *two or three.*

Two additional factors increase the complexity of this primary task. First, at every point during the task, one random bin is marked as unavailable and greyed out (with the disabled bin changed each time a block is placed in a bin). This forces users to remember all target bins. Second, to create a demanding auditory component to the primary task ensemble, the user hears a series of syllables playing in the task background, is given a target syllable to look out for, and is told that whenever they hear this syllable, the target bins and non-target bins are switched. In other words, the bins they should consider to place blocks in should be exchanged with those they were previously told to avoid. For example, if the user's target bins from among four bins are bins two and three, and they hear the target syllable, then future blocks will need to be placed instead into bins one and four. The syllables heard are selected from among

[4] These block colors were chosen for consistent visual processing, as blue is processed differently within the eye due to spatial and frequency differences of cones between red/green and blue. This did mean that our task was not accessible to red/green colorblind participants, requiring us to exclude data from colorblind participants.

(*bah*, *beh*, *boh*, *tah*, *teh*, *toh*, *kah*, *keh*, *koh*). These nine syllables were formed by combining three consonant sounds (b,t,k) with three vowel sounds (ah,eh,oh).

Secondary Task

As shown in Fig. 2, three times per experiment trial, the participant encounters a secondary task, in which the robot interrupts with a new request, asking the participant to move a particular, currently visible block, to a particular, currently accessible bin. Depending on experiment trial condition, this spoken request was sometimes accompanied by a MRDG. Unlike the long-term primary task that requires participants to remember the initial target block and keep track of the continuously changing target bins during the 90 second round, in the secondary task the robot asks participants to pick a different target block and place it in a different target bin, after which participants can continue the primary task.

3.2 Experimental Design

We used a Latin square counterbalanced within-subjects design with two within-subjects factors: Cognitive Load (4 loads) and Communication Style (3 styles).

Cognitive Load

Our first independent variable, cognitive load, was manipulated through our primary task. Following Beck and Lavie [27], we manipulated cognitive load by jointly manipulating memory constraints and target/distractor discriminability (cp. [26]), producing four load profiles: (1) all load considered low, (2) only working memory load considered high, (3) only visual perceptual load considered high, and (4) only auditory perceptual load considered high.

Working memory load was manipulated as follows: In the high working memory load condition, participants were required to remember the identities of three target bins out of a total of six visible bins, producing a total memory load of seven items: the three target bins, the target block color and shape, and the target syllable consonant and vowel. In all other conditions, participants were only required to remember the identities of two target bins out of a total of four visible bins, producing a total memory load of six items.

Visual perceptual load was manipulated as follows: In the high visual perceptual load condition, the target block was always difficult to discriminate from distractors due to sharing one common property with all distractors. For example, if the target block was a red cube, all distractors would be either red or cubes (but not both). In the low visual perceptual load condition, the target block was always easy to discriminate from distractors due to sharing no common properties with any distractors. For example, if the target block was a red cube, no distractors would be red or cubes.

Auditory perceptual load was manipulated as follows: In the high auditory perceptual load condition, the target syllable was always difficult to discriminate from distractors due to sharing one common property with all distractors.

For example, if the target syllable was *kah*, all distractors would either start with *k* or end with *ah* (but not both). In the low auditory perceptual load condition, the target syllable was always easy to discriminate from distractors due to sharing no common properties with any distractors. For example, if the target syllable was *kah*, no distractors would either start with *k* or end with *ah*.

Communication Style

Our second independent variable, communication style, was manipulated through our secondary task. Following Williams et al. [63, 64], we manipulated communication style by having the robot exhibit one of three behaviors:

1. During experiment blocks associated with the **complex language** communication style condition, the robot referred to objects using full referring expressions needed to disambiguate those objects (e.g., "the red sphere").
2. During experiment blocks associated with the **MR + complex language** communication style condition, the robot referred to objects using full referring expressions (e.g., "the red sphere"), paired with a MRDG (an arrow drawn over the red sphere).
3. During experiment blocks associated with the **MR + simple language** communication style condition, the robot referred to objects using minimal referring expressions (e.g., "that block"), paired with a MRDG (an arrow drawn over the object to which the robot was referring).

Following Williams et al. [63, 64], we did not examine the use of simple language without MR, which precludes referent disambiguation, resulting in the user needing to ask for clarification or guess between ambiguous options.

3.3 Measures

We expected performance improvements to manifest in our experiment in four ways: task accuracy, task response time, perceived mental workload, and perceived communicative effectiveness. These were measured as follows:

Accuracy was measured for both tasks by logging which object participants clicked on, determining whether this was the object intended by the task or by robot, and determining whether this object was placed in the correct bin.

Response time (RT) was measured for both primary and secondary tasks by logging time stamps at the moment participants interacted with virtual objects (both blocks and bins). In a primary task, whenever participants see a target block, their task is to pick-and-place it into any one of a particular set of bins. Thus, response time was measured as the delay between when the target block is first displayed and when the placement is completed because a new target block is immediately placed in a different location within the shown panel after a completed placement by the participant. In the secondary task, response time was measured as the time between the start of Pepper's utterance and the placement of the secondary target block.

Perceived mental workload was measured using a NASA Task Load Index (TLX) survey[21]. At the end of each experiment block, participants were asked to fill out a NASA TLX Likert 7-point scale survey across six categories: mental demand, physical demand, temporal demand, performance, effort, and frustration.

Perceived communicative effectiveness was measured using the modified Gesture Perception Scale [42] previously employed by Williams et al. [63, 64]: Participants were asked at the end of each experiment block to answer three 7-point Likert items on the effectiveness, helpfulness, and appropriateness of the robot's communication styles.

3.4 Procedure

Upon arriving at the lab, providing informed consent, and completing demographic and visual capability survey, participants were introduced to the task through both verbal instruction and an interactive tutorial.

The use of this interactive tutorial was motivated by several pilot tests that were run before conducting official trials. The initial pilot testers were given verbal instructions on how to use the HoloLens, how to complete their tasks in each round, and then were asked to start the 12 rounds. Feedback from these pilot tests showed that participants were not confident in their understanding of the HoloLens or game, so they struggled in the first couple rounds and improved with trial and error. This caused the participants' performance to be lower in the first few rounds than the later rounds, which made it hard to tell how the variations in the 12 rounds affect performance. To correct this, our team designed a tutorial scene that each participant completes at the start of the experiment, which further pilot studies demonstrated as addressing those concerns.

The tutorial scene walks the participant through a sample experimental round. When the participant starts the tutorial, they see a panel with text-instructions, a row of blocks, and four bins. Participants are walked through how to use the HoloLens air tap gesture to pick up blocks and put them in bins through descriptive text and an animation showing an example air tap gesture, and informed of task mechanics with respect to both target/non-target bins and temporarily disabled grey bins. Participants then start to hear syllables being played by the HoloLens. When the target syllable *teh* plays, the target and non-target bins switch. Each bin on screen is labeled as a 'target' or 'non-target', in order to help the participant understand what is happening when the target syllable plays. These labels are only shown in the tutorial and participants are reminded that they will have to memorize which bins are targets for the actual game. At the end of the tutorial the participant has to successfully put a target block in a target bin three times before they can start the experiment.

After completing this tutorial, participants engaged in each of the twelve (Latin square counterbalanced) experimental trials formed by combining the four cognitive load conditions and the three communication style conditions, with surveys administered after each experiment block. The length of the experiment, including surveys and breaks between each trial, was around 30 minutes.

3.5 Participants

36 participants were recruited from Colorado School of Mines (31 M, 5 F), ages 18-32. None had participated in previous studies from our laboratory.

3.6 Analysis

Data analysis was performed within a Bayesian framework using JASP 0.11.1 [48], using the default settings as justified by Wagenmakers et al. [51]. For each measure, a repeated measures analysis of variance (RM-ANOVA) [10, 33, 38] was performed, using communication style and cognitive load as random factors. Inclusion Bayes Factors across matched models ($\mathrm{BF}_{\mathrm{Incl}}$ [29]) were then computed for each candidate main effect and interaction. $\mathrm{BF}_{\mathrm{Incl}}$ for candidate effect E represents the ratio between two probabilities: the probability of our data being generated under models that included E, and the probability of our data being generated under models that did not include E. Therefore, this $\mathrm{BF}_{\mathrm{Incl}}$ represents the relative strength of evidence for an effect E, i.e.

$$\frac{\sum_{m \in M | e \in m} P(m|data)}{\sum_{m \in M | e \notin m} P(m|data)},$$

where e is an effect under consideration, and m is a candidate model in the space of candidate models M. When sufficient evidence was found for a main effect, the results were further analyzed using a post-hoc Bayesian t-test [24, 55] with a default Cauchy prior (center=0, r=$\frac{\sqrt{2}}{2}$=0.707).

The task accuracy was calculated as the ratio between the number of correct block placement and the total number of block placements, with 0 being complete failure and 1 being correct placement for each placed block within a trial.

Finally, transformations were applied to response time data. Since response time distributions are often not Gaussian (normally distributed) but rather have a long right-tail, logarithmic *log(RT)* transformations are often used by researchers to handle such data [56].

A Shapiro-Wilk test of normality indicated (p<.01) that data in all conditions was non-normally distributed. While an assumption of normal distribution is not necessary for our analyses due to our use of a Bayesian analysis framework, the reason for non-normality in our data was asymmetry, with a long right tail and a number of extreme positive outliers. These considerations together suggested the need for data transformation, regardless of analysis framework. To reduce sensitivity to non-normally distributed outliers and induce a more normal data distribution, we applied a log transformation on all response time data.

4 Results

4.1 Response Time

Strong evidence was found against any effect of communication style or imposed cognitive load on primary task response time, with all $\mathrm{BFs}_{\mathrm{Incl}} < 0.028$ for an effect. These results fail to support either hypothesis, with no benefit of MRDGs observed in *any* condition. Our results provided strong evidence for an effect of communication style ($\mathrm{BF_{Incl}}$=17.860) on secondary task response time, as shown in Fig. 3, but evidence against an effect of imposed workload ($\mathrm{BF_{Incl}}$=0.017), or of an interaction between workload and communication style, on response time ($\mathrm{BF_{Incl}}$=0.018). A post-hoc Bayesian t-test analyzing the effect of communication style revealed extreme evidence (BF=601.460) for a difference in response time between the complex language condition (μ = 2.095, σ = 0.325; untransformed μ = 8.877 seconds, σ = 4.072 seconds) and the MR + complex language condition (μ = 1.955, σ = 0.323; untransformed μ = 7.779, σ = 3.877), weak evidence (BF=1.551) for a difference in response time between the complex language condition and MR + simple language condition (μ = 2.006, σ = 0.436; untransformed μ = 8.764, σ = 6.203), and moderate evidence (BF=0.203) *against* a difference between the MR + complex language and MR + simple language conditions. In other words, when you use MR, language makes little difference. And when you use complex language only, having MR is a big help. The evidence against an effect of workload but for an effect of communication style provides partial support for the *Universal Benefit Hypothesis*, as MRDGs do indeed provide task-oriented benefits regardless of level and type of cognitive load, but only when paired with rich referring expressions.

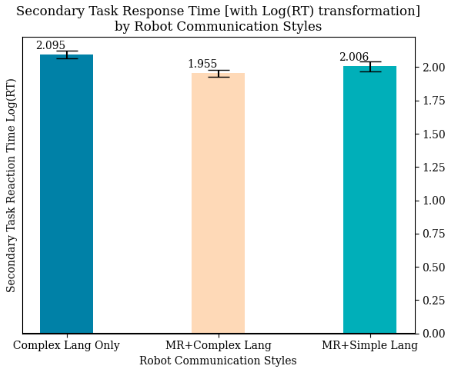

Fig. 3. Effect of communication styles on participant's secondary task log(RT). Error bars represent standard errors.

4.2 Accuracy

Strong evidence was found *against* effects of communication style or imposed cognitive load on primary or secondary task accuracy (All $\mathrm{BFs_{Incl}} < 0.033$ for an effect). Mean primary task accuracy was 0.706 (σ = 0.261). Mean secondary task accuracy was 0.984 (σ = 0.074). These results fail to support either hypothesis, with no benefit of MRDGs observed in *any* condition.

4.3 Perceived Mental Workload

Strong evidence was found *against* any effects of communication style or imposed cognitive load on perceived mental workload ($\mathrm{BF}_{\mathrm{Incl}}$ between 0.006 and 0.040 for an effect). Aggregating across conditions, TLX score sums had a mean of 21.109 out of 42 points ($\sigma = 5.443$). Thus, most participants' perceived workload data was almost perfectly centered around "medium load". These results fail to support either hypothesis, with no benefit of MRDGs observed in *any* condition.

4.4 Perceived Communicative Effectiveness

Anecdotal to strong evidence was found *against* any effects of communication style or cognitive load on perceived communicative effectiveness ($\mathrm{BF}_{\mathrm{Incl}}$ between 0.049 and 0.117 for an effect on all questions). Participants' perceived communicative effectiveness had a mean of 5.611 out of 7 ($\sigma = 1.208$). These results fail to support either hypothesis, with no benefits observed in *any* condition.

5 Discussion

Our results provide partial support for the *Universal Benefit Hypothesis*: while the types of task-oriented benefits of MRDGs previously observed in some recent laboratory studies [18, 7] were largely unobserved, these benefits *were* observed, *regardless* of cognitive load, for secondary task response time; but *only* when MRDGs were paired with complex language. These results suggest that the primary benefit of MRDGs in robot communication lies in their ability to increase users' speed at performing a secondary task by reducing the time taken to perform constituent visual searches (especially when paired with complex referring expressions), regardless of the level and type of workload users are experiencing.

Moreover, our results have interesting (albeit non-identical) parallels with previous work *not* performed in realistic task environments [64], which found that participants demonstrated slower response times when complex language alone was used, with no clear differences between simple and complex language when pairing language with MRDGs. That previous study also suggested that people found a robot to be more likable when it used longer more natural referring expressions. When combined with the results of our own experiment, this suggests that robots can likely pair complex referring expression with mixed reality gestures without worrying about cognitively overloading their interlocutors.

Our results overall provide evidence against the four Cognitive Contextual Benefit Hypotheses, casting doubt on the potential of adaptive automation to provide benefits in mixed reality human robot dialogue. While this hypothesis would have predicted that the differences between communication styles under different cognitive load profiles would primarily be grounded in whether communication style was overall visual or overall auditory, in fact what we observed is that visual augmentations, especially when paired with complex referring expressions, may *always* be helpful for a secondary task (when paired with complex language), regardless of level and type of imposed workload.

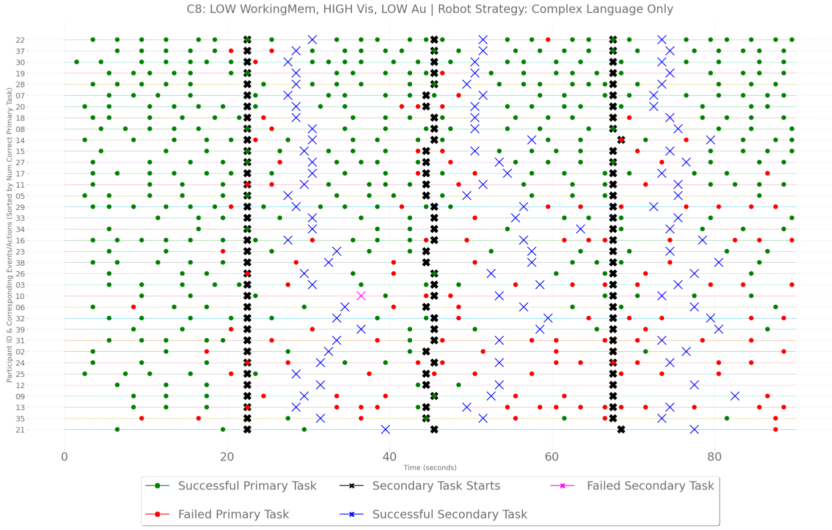

Fig. 4. Visualization of participant performance in the *Complex Language Only* / *High visual perceptual load* condition. Each row depicts the performance of one participant, with participants presented in decreasing order of primary task accuracy. Green and red dots represent primary task block placement times, with green dots indicating successful placement and red dots indicating unsuccessful placement (i.e., placement involving an incorrect block or incorrect bin). Xes represent secondary task instruction and completion times, with dark Xes indicating when the robot started uttering a secondary task request, blue Xes indicating when participants successfully completed that secondary task, and pink Xes indicating when participants failed a secondary task (i.e., placement involving an incorrect block or incorrect bin).

Similarly, we found no effect of imposed workload or gesture on perceived workload or perceived effectiveness. This may have been for at least three reasons. First, secondary response time differences might simply not have been large enough for participants to notice: the observed differences were on the order of one second of response time when overall secondary task response time was around 7.5 seconds, representing only a 15% secondary task efficiency increase.

Second, the benefits of mixed reality were only seen on speed of response to the relatively rare secondary tasks, and not to the much more frequent primary tasks. Participants may have primarily – or only – considered their primary task when reporting their perceived workload and perceived effectiveness.

Finally, while participants' TLX scores had a mean of 21.109 out of 42 points in all conditions (i.e., the data was nearly perfectly centered around "medium" load), analysis of individual performance trajectories demonstrates that the task was sufficiently difficult that many participants experienced catastrophic primary task shedding. Consider Fig. 4, which shows the results of each participant within one of the twelve conditions, with participants listed in decreasing order of primary task accuracy. As shown in this figure, and described in the caption, the bottom 50% of participants experienced large numbers of failures, with many of these participants experiencing a failure that they never recovered from immediately after a secondary task, perhaps due to missing an auditory cue during that secondary task. While Fig. 4 shows only one condition (the *complex language/High visual load* condition) for the sake of space, in fact all twelve condition plots show similar results.

This suggests it may be too early to cast doubt on the potential use of adaptive automation in human-robot dialogue, as our results may have been due to subtle aspects of our experimental or task design rather than universal principles of human cognition. Moreover, our experiment had a number of limitations that further motivate the need for future work.

5.1 Limitations and Future Work

While our study provides evidence of the effect of MRDGs on human's task response time, it has key limitations to address in future work. Given the catastrophic errors experienced by some participants, and given that all twelve condition plots show similar results, our experimental setup should be reconsidered. For example, some participants failed early into the game and completely lost track of what block to place in what bin. Providing real-time, directive cues might help participant recover from errors. However, the purpose of a challenging primary task is to impose high workload on the participants and to observe how different communication styles can help enhance human task performance under cognitive overload. The poor performance observed in our experiment demonstrates the effect of the cognitive demanding primary task, but the catastrophic primary task shredding complicated our effort to unravel the impact on accuracy, response time, and perceived workload. Additional consideration is needed to design ways that recovery hints can be presented (visual or auditory) without interfering with the imposed workload profiles during the experiment.

Additionally, we received feedback from some participants during the debriefing that they felt the series of syllables playing in the task background (e.g., *bah*, *beh*, *boh*, *tah*, *teh*, *toh*, *kah*, *keh*, *koh*) could easily be misheard. After missing the auditory cue that signals the switch of the target and non-target bins, they started to guess the target bins to attempt to proceed with the primary task. We recommend in future research to use distinguishable sounds instead of these syllables in order to improve auditory discrimination.

Another direction for future research is to use eye- or hand-tracking (e.g., through the HoloLens 2) to more precisely capture response time. For example, it would have been advantageous to capture the delay between when a target block first appeared and when participants first gazed at it, or the time between the block's first appearance and the participants movement of their hands to commit to a new target goal. Rather than measuring response time as *TimeBlockPlaced* – *TimeBlockAppeared* in this experiment, researchers could use hand trajectories and movement data to infer the underlying cognitive processes, such as mental processing time and midflight corrections (i.e., when participants initially move their hands towards an incorrect block, and then perhaps under the suggestion of the robot, switch their target to the correct block).

Furthermore, devices such as HoloLens 2 enable new input modalities that allow completely natural hand gestures rather than the simple gaze-and-commit (e.g., air tap) interaction of the Hololens 1. In our experiment, participants were given time and tutorials to become acquainted with the headset and practice the air-tap hand gesture. Even though most participants expressed that they felt comfortable with the headset and interaction to start the experiment, some still struggled to pick-and-place the virtual blocks, affecting the measurement of response time, interfering with the load placed by the primary task. Accordingly, the system's limitations led to issues in differentiation between the delay participants took to figure out how to use the gesture vs. the true delay caused by the cognitively taxing primary task.

Another limitation of our experiment was the number of participants recruited. We were able to recruit 36 participants pre-COVID 19, and while our current analysis provided evidence against effects of workload profiles on task time, a larger participant pool would have allowed for more decisive conclusions.

6 Conclusion

We examined the effectiveness of different combinations of natural language reference and MRDG under different types of mental workload, through a 36-participant Mixed Reality Robotics experiment. We found that, for our verbal and nonverbal communication strategies and workload manipulations, MRDGs improve the effectiveness of users by shortening their response time in a secondary visual search tasks, regardless of underlying level and type of cognitive load, providing partial support for a *Universal Benefit Hypothesis*. Moreover, we found this to be especially true when MRDGs were paired with complex referring expressions rather than concise demonstrative pronouns. These results will help inform future efforts in mixed reality robot communication by demonstrating how MRDGs and natural language referring expression should be paired to best enhance the effectiveness of robots' human teammates.

Acknowledgement

This research was funded in part by NSF grants IIS-1909864 and CNS-1810631.

References

1. Amor, H.B., Ganesan, R.K., Rathore, Y., Ross, H.: Intention projection for human-robot collaboration with mixed reality cues. In: Int'l WS on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI) (2018)
2. Andersen, R.S., Madsen, O., Moeslund, T.B., Amor, H.B.: Projecting robot intentions into human environments. In: Int'l Symposium on Robot and Human Interactive Communication (RO-MAN). pp. 294–301 (2016)

3. Azuma, R.: A survey of augmented reality. Presence: Teleoperators & Virtual Environments **6**, 355–385 (1997)
4. Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., MacIntyre, B.: Recent advances in augmented reality. IEEE computer graphics and applications (2001)
5. Billinghurst, M., Clark, A., Lee, G., et al.: A survey of augmented reality. Foundations and Trends® in Human–Computer Interaction **8**(2-3), 73–272 (2015)
6. Meyer zu Borgsen, S., Renner, P., Lier, F., Pfeiffer, T., Wachsmuth, S.: Improving human-robot handover research by mixed reality techniques. In: Int'l WS on Virtual, Aug. and Mixed Reality for Human-Robot Interaction (VAM-HRI) (2018)
7. Brown, L., Hamilton, J., Han, Z., Phan, A., Phung, T., Hansen, E., Tran, N., Williams, T.: Best of both worlds? combining different forms of mixed reality deictic gestures. ACM Transactions on Human-Robot Interaction (2022)
8. Chakraborti, T., Sreedharan, S., Kulkarni, A., Kambhampati, S.: Alternative modes of interaction in proximal human-in-the-loop operation of robots. arXiv preprint arXiv:1703.08930 (2017)
9. Cheli, M., Sinapov, J., Danahy, E.E., Rogers, C.: Towards an augmented reality framework for k-12 robotics education. In: Int'l WS on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI) (2018)
10. Crowder, M.J.: Analysis of repeated measures. Routledge (2017)
11. Dudley, A., Chakraborti, T., Kambhampati, S.: v2v communication for augmenting reality enabled smart huds to increase situational awareness of drivers (2018)
12. Frank, J.A., Moorhead, M., Kapila, V.: Mobile mixed-reality interfaces that enhance human–robot interaction in shared spaces. Front. Rob. & AI **4**, 20 (2017)
13. Ganesan, R.K., Rathore, Y.K., Ross, H.M., Amor, H.B.: Better teaming through visual cues: how projecting imagery in a workspace can improve human-robot collaboration. IEEE Robotics & Automation Magazine **25**(2), 59–71 (2018)
14. Goktan, I., Ly, K., Groechel, T.R., Mataric, M.: Augmented reality appendages for robots: Design considerations and recommendations for maximizing social and functional perception. In: Int'l WS on Virt., Aug., and Mixed Real. for HRI (2022)
15. Green, S.A., Billinghurst, M., Chen, X., Chase, J.G.: Human-robot collaboration: A literature review and augmented reality approach in design. Int'l journal of advanced robotic systems **5**(1), 1 (2008)
16. Groechel, T., Shi, Z., Pakkar, R., Matarić, M.J.: Using socially expressive mixed reality arms for enhancing low-expressivity robots. In: Int'l Conference on Robot and Human Interactive Communication (RO-MAN). pp. 1–8. IEEE (2019)
17. Groechel, T.R., Walker, M.E., Chang, C.T., Rosen, E., Forde, J.Z.: Tokcs: Tool for organizing key characteristics of vam-hri systems. Rob. & Autom. Mag. (2021)
18. Hamilton, J., Phung, T., Tran, N., Williams, T.: What's the point? tradeoffs between effectiveness and social perception when using mixed reality to enhance gesturally limited robots. In: Proc. HRI (2021)
19. Hamilton, J., Tran, N., Williams, T.: Tradeoffs between effectiveness and social perception when using mixed reality to supplement gesturally limited robots. In: Int'l WS on Virtual, Augmented, and Mixed Reality for HRI (2020)
20. Han, Z., Zhu, Y., Phan, A., Garza, F.S., Castro, A., Williams, T.: Crossing reality: Comparing physical and virtual robot deixis. In: Int'l Conf. HRI (2023)
21. Hart, S., Staveland, L.: Development of NASA-TLX (Task Load Index): Results of empirical and theorical research, pp. pp 139 – 183. Amsterdam (1988)
22. Hedayati, H., Walker, M., Szafir, D.: Improving collocated robot teleoperation with augmented reality. In: Int'l Conference on Human-Robot Interaction (2018)

23. Hirshfield, L., Williams, T., Sommer, N., Grant, T., Gursoy, S.V.: Workload-driven modulation of mixed-reality robot-human communication. In: ICMI WS on Modeling Cognitive Processes from Multimodal Data. p. 3. ACM (2018)
24. Jeffreys, H.: Significance tests when several degrees of freedom arise simultaneously. Proc. Royal Society of London. Series A, Math. and Phys. Sci. (1938)
25. Kahneman, D.: Attention and effort (1973)
26. Lavie, N.: Perceptual load as a necessary condition for selective attention. Journal of Experimental Psych.: Human perception and performance **21**(3), 451 (1995)
27. Lavie, N.: The role of perceptual load in visual awareness. Brain research (2006)
28. MacDonald, W.: The impact of job demands and workload on stress and fatigue. Australian Psychologist **38**(2), 102–117 (2003)
29. Mathôt, S.: Bayes like a baws: Interpreting bayesian repeated measures in JASP [blog post]. cogsci.nl/blog/interpreting-bayesian-repeated-measures-in-jasp (2017)
30. Matuszek, C., Bo, L., Zettlemoyer, L., Fox, D.: Learning from unscripted deictic gesture and language for human-robot interactions. In: AAAI (2014)
31. Mavridis, N.: A review of verbal and non-verbal human–robot interactive communication. Robotics and Autonomous Systems **63**, 22–35 (2015)
32. Milgram, P., Zhai, S., Drascic, D., Grodski, J.: Applications of augmented reality for human-robot communication. In: Int'l Conf. Intel. Robots and Systems (1993)
33. Morey, R., Rouder, J.: Bayesfactor (version 0.9. 9) (2014)
34. Navon, D., Gopher, D.: On the economy of the human-processing system. Psychological review **86**(3), 214 (1979)
35. Norman, D.A., Bobrow, D.G.: On data-limited and resource-limited processes. Cognitive psychology **7**(1), 44–64 (1975)
36. Peters, C., Yang, F., Saikia, H., Li, C., Skantze, G.: Towards the use of mixed reality for hri design via virtual robots. In: Int'l WS on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI) (2018)
37. Rosen, E., Whitney, D., Phillips, E., Chien, G., Tompkin, J., Konidaris, G., Tellex, S.: Communicating robot arm motion intent through mixed reality head-mounted displays. In: Robotics Research, pp. 301–316. Springer (2020)
38. Rouder, J.N., Morey, R.D., Speckman, P.L., Province, J.M.: Default bayes factors for anova designs. Journal of Mathematical Psychology **56**(5), 356–374 (2012)
39. Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., Joublin, F.: To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. Int'l Journal of Social Robotics **5**(3), 313–323 (2013)
40. Salem, M., Kopp, S., Wachsmuth, I., Rohlfing, K., Joublin, F.: Generation and evaluation of communicative robot gesture. Int'l Journal of Social Rob. **4**(2) (2012)
41. Sanders, A.: Dual task performance (2001)
42. Sauppé, A., Mutlu, B.: Robot deictics: How gesture and context shape referential communication. In: Int'l Conference on Human-Robot Interaction (HRI) (2014)
43. Schönheits, M., Krebs, F.: Embedding ar in industrial hri applications. In: Int'l WS on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI) (2018)
44. Sibirtseva, E., Kontogiorgos, D., Nykvist, O., Karaoguz, H., Leite, I., Gustafson, J., Kragic, D.: A comparison of visualisation methods for disambiguating verbal requests in human-robot interaction. In: Int'l Sym. Rob. Hum. Inter. Comm. (2018)
45. Siéroff, E.: Attention: Multiple resources (2001)
46. Sportillo, D., Paljic, A., Ojeda, L., Partipilo, G., Fuchs, P., Roussarie, V.: Learn how to operate semi-autonomous vehicles with extended reality (2018)
47. Szafir, D.: Mediating human-robot interactions with virtual, augmented, and mixed reality. In: Int'l Conference on Human-Computer Interaction (2019)

48. Team, J.: Jasp (version 0.8.5.1)[computer software] (2018)
49. Tellex, S., Gopalan, N., Kress-Gazit, H., Matuszek, C.: Robots that use language. Annual Review of Control, Robotics, and Autonomous Systems **3** (2020)
50. Van Krevelen, D., Poelman, R.: A survey of augmented reality technologies, applications and limitations. Int'l journal of virtual reality **9**(2), 1–20 (2010)
51. Wagenmakers, E., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J.: Bayesian inference for psychology, Part II: Example applications with JASP. Psychonomic Bulletin and Review **25**(1), 35–57 (2018)
52. Walker, M., Hedayati, H., Lee, J., Szafir, D.: Communicating robot motion intent with augmented reality. In: Int'l Conference on Human-Robot Interaction (2018)
53. Walker, M., Phung, T., Chakraborti, T., Williams, T., Szafir, D.: Virtual, augmented, and mixed reality for human-robot interaction: A survey and virtual design element taxonomy. https://arxiv.org/abs/2202.11249 (2022)
54. Weng, T., Perlmutter, L., Nikolaidis, S., Srinivasa, S., Cakmak, M.: Robot object referencing through legible situated projections. In: Int'l Conference on Robotics and Automation (ICRA) (2019)
55. Westfall, P.H., Johnson, W.O., Utts, J.M.: A bayesian perspective on the bonferroni adjustment. Biometrika **84**(2), 419–427 (1997)
56. Whelan, R.: Effective analysis of reaction time data. The Psychological Record **58**(3), 475–482 (2008)
57. Wickens, C.D.: Processing resources and attention. Multiple-task perf. (1991)
58. Wickens, C.D.: Multiple resources and performance prediction. Theoretical issues in ergonomics science **3**(2), 159–177 (2002)
59. Wickens, C.D.: Multiple resources and mental workload. Hum. fac. **50**(3) (2008)
60. Wickens, C.D., Santamaria, A., Sebok, A.: A computational model of task overload management and task switching. In: Human factors and ergonomics society annual meeting. vol. 57, pp. 763–767. SAGE Publications Sage CA: Los Angeles, CA (2013)
61. Wickens, C.D., Tsang, P.: Handbook of human-systems integration. APA (2014)
62. Wickens, C.D., Vidulich, M., Sandry-Garza, D.: Principles of scr compatibility with spatial and verbal tasks: The role of display-control location and voice-interactive display-control interfacing. Human factors **26**(5), 533–543 (1984)
63. Williams, T., Bussing, M., Cabrol, S., Boyle, E., Tran, N.: Mixed reality deictic gesture for multi-modal robot communication. In: Int'l Conf. HRI (2019)
64. Williams, T., Bussing, M., Cabrol, S., Lau, I., Boyle, E., Tran, N.: Investigating the potential effectiveness of allocentric mixed reality deictic gesture. In: Int'l Conference on Virtual, Augmented, and Mixed Reality (2019)
65. Williams, T., Szafir, D., Chakraborti, T.: The reality-virtuality interaction cube. In: Int'l WS on Virtual, Augmented, and Mixed Reality for HRI (2019)
66. Williams, T., Szafir, D., Chakraborti, T., Ben Amor, H.: Virtual, augmented, and mixed reality for human-robot interaction. In: Int'l Conference on Human-Robot Interaction (LBRs). pp. 403–404. ACM (2018)
67. Williams, T., Tran, N., Rands, J., Dantam, N.T.: Augmented, mixed, and virtual reality enabling of robot deixis. In: Int'l Conference on Virtual, Augmented and Mixed Reality. pp. 257–275. Springer (2018)
68. Williams, T., Yazdani, F., Suresh, P., Scheutz, M., Beetz, M.: Dempster-shafer theoretic resolution of referential ambiguity. Autonomous Robots **43**(2) (2019)
69. Zhou, F., Duh, H.B.L., Billinghurst, M.: Trends in augmented reality tracking, interaction and display: A review of ten years of ismar. In: Int'l Symposium on Mixed and Augmented Reality. pp. 193–202. IEEE (2008)