# Deictic Adaptation in a Virtual Environment

Nikhil Krishnaswamy and James Pustejovsky
Brandeis University

Spatial Cognition 2018
Tübingen, Germany
September 7, 2018

# Introduction

- We examine the role of deixis in peer-to-peer communication between humans and computers
- Deixis is denotative within a situated space
- How humans use deixis relates their spatial model of the environment
- Interaction with computers (i.e., in virtual environments) is inherently different from real-world environments
- We examine how users adapt their use of deixis in a virtual environment under different experimental conditions in the course of a collaboration with a computer agent

# Introduction

- In human interactions, assumptions about the interlocutor influence communication style, message design, available vocabulary and expression modality (Edwards and Shepherd, 2004; Arbib, 2008)
- When collaborating agents each have incomplete knowledge of a situation, they rely on their interlocutor(s) to clarify or provide instructions, facilitated by
  - imagining situation from a different perspective (Bergen, 2012)
  - neural structures (e.g., mirror neurons) (Arbib and Rizzolatti, 1996)

# Related Work

- Two agents jointly experiencing a localized event are *co-situated* and *co-perceptive*
- Collaborating agents *co-intend* to the task and *co-attend* to the situation
- These parameters come together in a theory of *common ground* (Clark, Schreuder, and Buttrick, 1983; Stalnaker, 2002; Asher and Gillies, 2003; Pustejovsky, 2018)
  - Rich, diverse literature on common ground exists (e.g., Clark and Brennan, 1991; Stalnaker, 2002; Tomasello and Carpenter, 2007)

# Related Work

- Some problems in a strictly presuppositional view of common ground (e.g., Abbott, 2008)
- Mitigated by mechanisms such as "accommodation" (cf. Lewis, 1979)
- When the assumptions that facilitate these mechanisms are not in force, common ground breaks down
  - Common ground between a human and an animal is limited (Kirchhofer et al., 2012)
  - Common ground between human and computer/robot is also limited
  - No accommodation mechanism exists in a computer system unless put there by developers

# Related Work

- Unlike an animal, computational agents are built to approximate (subset of) human behavior
- As computational agents become more sophisticated, users expect them to behave more like humans (David et al., 2006; Fussell et al., 2008)

# Mental Simulation and Mind Reading

- Mental Simulations
  Graesser et al (1994), Barselou (1999), Zwaan and Radvansky (1998), Zwaan and Pecher (2012)
- Embodiment:
  Johnson (1987), Lakoff (1987), Varela et al. (1991), Clark (1997), Lakoff and Johnson (1999), Gibbs (2005)
- Mirror Neuron Hypothesis:
  Rizzolatti and Fadiga (1999), Rizzolatti and Arbib (1998), Arbib (2004)
- Simulation Semantics
  Goldman (1989), Feldman et al (2003), Goldman (2006), Feldman (2010), Bergen (2012), Evans (2013)

# Communication in Virtual Environments

- **How does the expectation of near-human capability, plus the agent's lack of sophisticated pragmatic mechanisms, manifest where some understanding of common ground is required to complete a task?**
- We previously examined factors in computational common ground (Pustejovsky et al., 2017), continued here
- We integrate multimodal model of semantics (Pustejovsky and Krishnaswamy, 2016; Krishnaswamy and Pustejovsky, 2016a) with a realtime gesture recognition (Wang et al., 2017b).
- Human communicated spatially-grounded instructions in a collaborative task (Krishnaswamy et al., 2017; Narayana et al., 2018)
- How do human users adapt their deictic techniques based on variant spatial cues?

# Communication in Virtual Environments

- **Deixis!**
  - A basic spatially-grounded gesture
  - A general mode of reference that refers to an orientation, location, or object inside it (cf. Ballard et al., 1997)
- Object indicated by deixis is usually current focus (Brooks and Breazeal, 2006)
- Mismatch in frame of reference or known information may lead to confusion about object or coordinate indicated by deixis (Hindmarsh et al., 2000; Williams and Scheutz, 2017)
- Speed of pointing inversely correlates to the difficulty of the pointing task being performed (Papaxanthis, Pozzo, and Schieppati, 2003; Zhai, Kong, and Ren, 2004)

# Deixis in Virtual Environments

$$
\left[\begin{array}{l}
\textbf{point} \\
\text{LEX} = \left[\begin{array}{l} \text{PRED} = \textbf{point} \\ \text{TYPE} = \textbf{assignment} \end{array}\right] \\
\text{TYPE} = \left[\begin{array}{l}
\text{HEAD} = \textbf{assignment} \\
\text{ARGS} = \left[\begin{array}{l} \text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:finger} \\ \text{A}_3 = \textbf{z:location} \\ \text{A}_4 = \textbf{w:physobj•location} \end{array}\right] \\
\text{BODY} = \left[\begin{array}{l} \text{E}_1 = \mathit{extend}(x, y) \\ \text{E}_2 = \mathit{def}\left(\mathit{vec}(x \rightarrow y \times z), \mathit{as}(w)\right) \end{array}\right]
\end{array}\right]
\end{array}\right]
$$

Figure: VoxML semantics (Pustejovsky and Krishnaswamy, 2016) for a [[POINT]] gesture. $\text{A}_4$, $w$, shows the compound typing (a la Generative Lexicon (Pustejovsky, 1995)) of the indicated region and objects within that region.

# Experimental Platform

- Multimodal human-computer interaction
- Gesture (Wang et al., 2017a) and natural language in a 3D simulated environment, created with VoxML platform and VoxSim (Krishnaswamy and Pustejovsky, 2016a; Krishnaswamy and Pustejovsky, 2016b)
- Real time gesture recognition (Microsoft Kinect depth data on ResNet-style DCNNs)

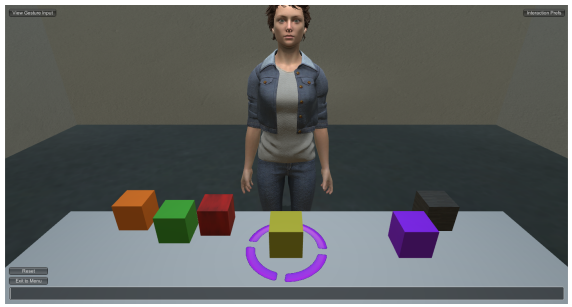


Figure: VoxSim Environment

# Experimental Setup

- Based on human-to-human elicitation studies (Wang et al., 2017a)
  - "Signaler" has target structure
  - Must instruct "builder" to build it
  - Both people situated before a table, connected by video feed, only builder has blocks
  - Table began to serve as point of reference, influenced creation of gesture recognition system
- Mirroring exercise: $Point_G \rightarrow Loc \mid Obj \quad Point_G \rightarrow Loc\prime \mid Obj\prime$ from signaler's table space to the builder's table space
- Without common reference point (e.g., table), studies show subjects default to pointing relative to other context
  - Free-floating point within VR environment (Wraga, Creem-Regehr, and Proffitt, 2004)
  - Screen display (Hindmarsh and Heath, 2000; Moeslund, Störring, and Granum, 2001)

# Experimental Setup

Krishnaswamy and Pustejovsky, 2018

- System requirements for deixis conflict with users' documented tendencies
- Creates opportunity to study if and how users adapt deixis to the system
- Users collaborated with avatar to build test pattern: 3-step, 6-block staircase

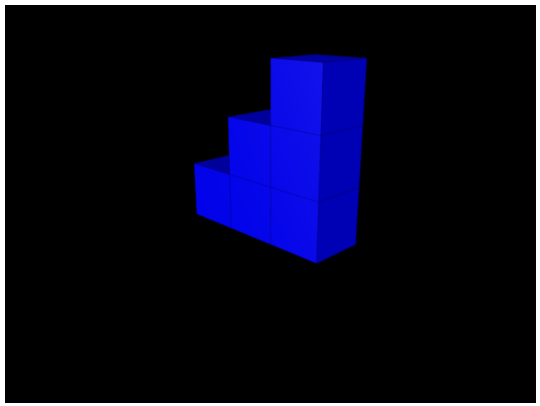
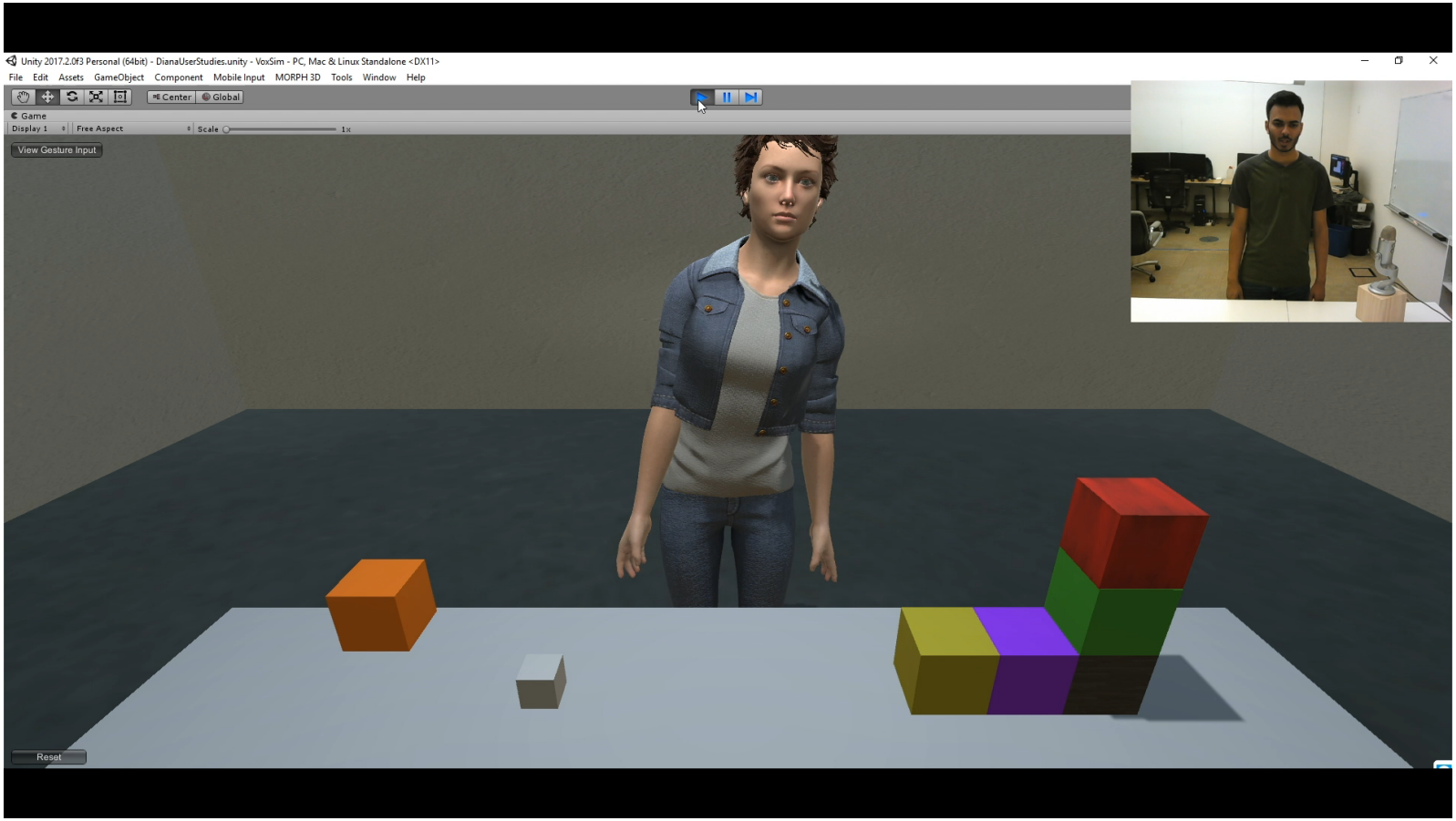


Figure: Test pattern given to naive users

# Experiment: Video Demo




▸ Link

# Experimental Setup

Krishnaswamy and Pustejovsky, 2018

- 20 CS grad students
- No knowledge of the system or gesture vocabulary
- 10 with table, 10 without

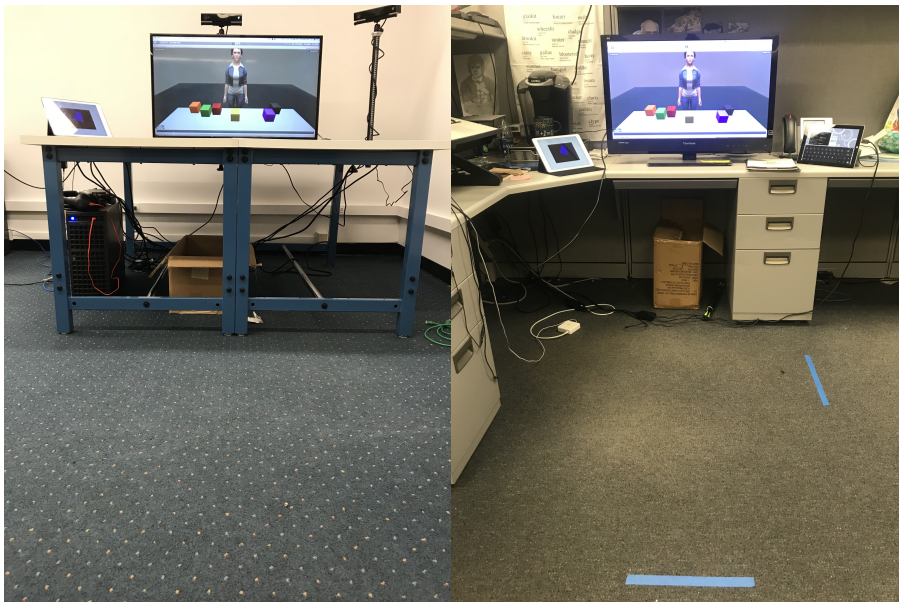


Figure: Variant environmental setups

# Experimental Setup

Krishnaswamy and Pustejovsky, 2018

- Each environment divided in two conditions:

| Condition | Physical Table | Supplemental Information |
|---|---|---|
| 1 | present (**A**) | none |
| 2 | absent (**B**) | none |
| 3 | present (**A**) | Physical table extends virtual table |
| 4 | absent (**B**) | Virtual table extends into real world |

Table: 5 subjects were placed in each experimental condition

- Supplemental info served as an implicit "hint" that table/imagined table space had role to play

# Experimental Setup

- Log format: `INDEX, SYMBOL, CONTENT, TIMESTAMP`
- Here, focusing only on human pointing gestures (`HP`).

```
1 HG engage start 1.145281
2 AS "Hello." 1.145281
3 HP r,-0.25,-0.87 4.889832
4 HP r,-0.16,-1.21 4.928307
5 HP r,-0.07,-1.18 4.960413
6 HP r,-0.03,-1.06 5.040221
7 HP r,-0.09,-0.95 5.072867
8 HP r,-0.07,-0.27 5.15642
...
73 HP r,-0.08,11.69 8.552608
74 HG right point high,-0.02,5.45 8.588802
75 AS "Are you pointing here?" 8.588802
```

- *Successful pointing*: Point sequence, avatar response, followed by positive acknowledgment
- *Failed pointing*: Point sequence, avatar response, followed by negative acknowledgment

# Experimental Analysis

- Time to successfully point: interval from start of pointing (move #3 in example) to recognition of location (move #74 in example)
  - Only in blocks where pointing event precedes positive acknowledgment
- If user adapts deictic strategy to the system, times to complete a successful pointing should decrease as user proceeds further into the interaction
- Adaptation in pointing times modeled as a *learning rate* (Wright, 1936)
- Examine in which conditions users adapt a strategy more quickly

# Preprocessing

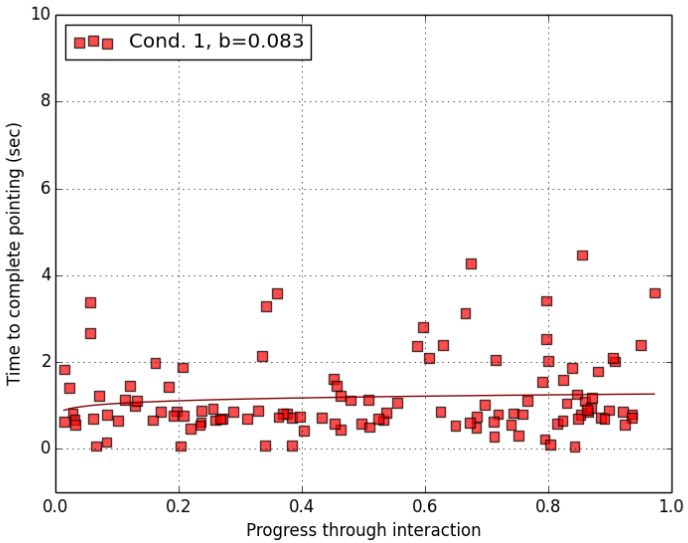
- Aggregated the data from all sessions of all users in a single condition
- Removed outliers (times lying outside the interquartile range for the distribution of all times logged, independent of condition)
- Sessions all of different lengths, so we cannot use raw duration of an interaction as the independent variable
  - Normalized by plotting a user's pointing times against the *percentage* of the total interaction completed to that point

# Results

- Plotted data in two ways:
  - Raw times taken to complete successful pointing events against percentage of interaction completed
    - Assess a learning curve (as a power law: $y_n = ax^{b_\rho}$) for an average user in a given condition
    - Does raw time to successfully complete a pointing over the course of an interaction decline, stay flat, or increase?
  - *Ratio* between time to complete successful pointing event and user's *geometric mean* time to complete a successful pointing, against the percentage of interaction completed.
    - Users may have different "natural aptitudes" with the system
    - Normalizes some of the variation due to given subject's "set point"
    - Using geometric mean allows linear regression plot ($log\ y_n = log\ a + b_\mu\ log\ x$), and more intuitive representation

# Results

- X: % progress through trial; Y (L): time to complete successful pointing; Y (R): time to complete successful pointing, as ratio to the geometric mean of all user's recorded pointing times
- Best fit line is shown as a least-squares fitted power law (L), and a linear regression (R)



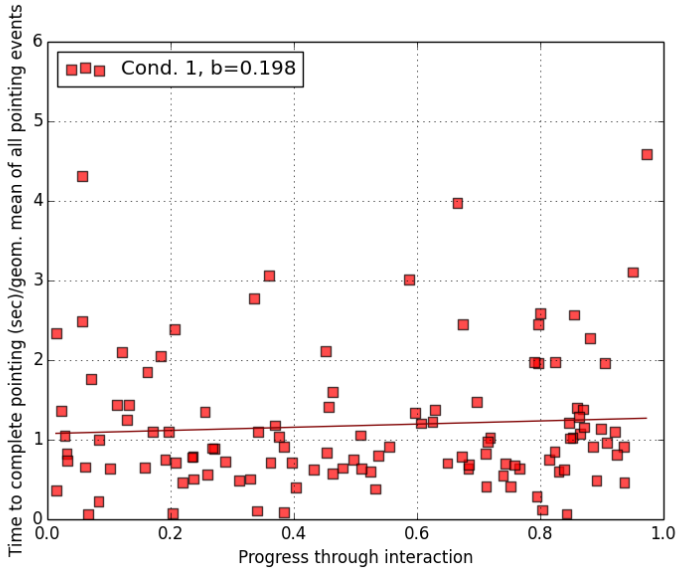
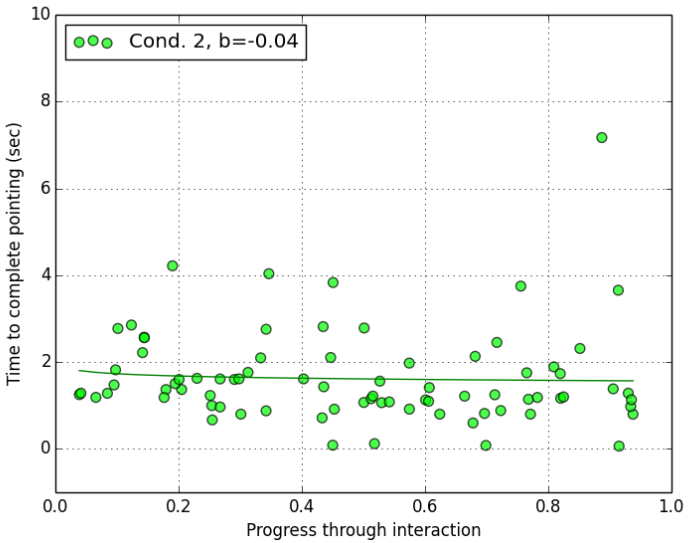




Figure: Results with table, no hint. $b_\rho \approx 0.083$, $s \approx 1.059$; $b_\mu \approx 0.198$

# Results

- X: % progress through trial; Y (L): time to complete successful pointing; Y (R): time to complete successful pointing, as ratio to the geometric mean of all user's recorded pointing times
- Best fit line is shown as a least-squares fitted power law (L), and a linear regression (R)



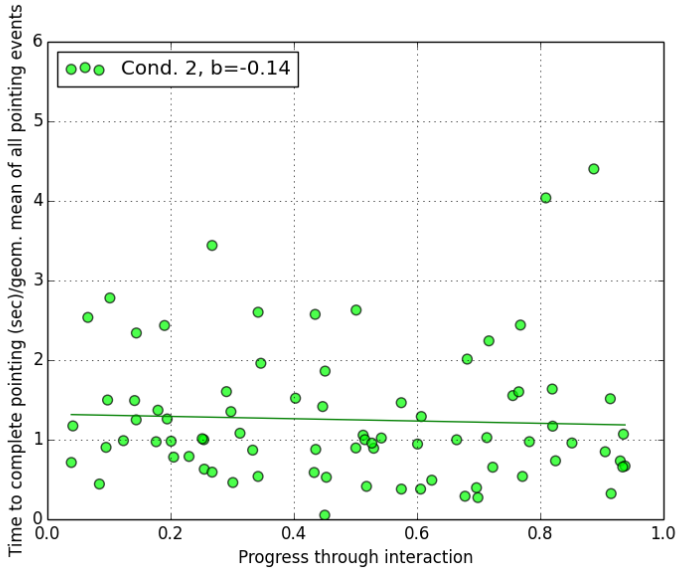
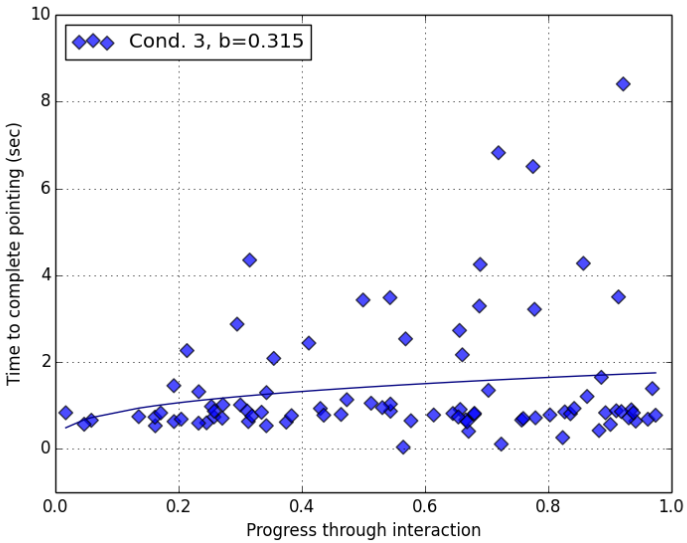




Figure: Results without table, no hint. $b_\rho \approx$ -0.044, $s \approx 0.970$; $b_\mu \approx$ -0.144

# Results

- X: % progress through trial; Y (L): time to complete successful pointing; Y (R): time to complete successful pointing, as ratio to the geometric mean of all user's recorded pointing times
- Best fit line is shown as a least-squares fitted power law (L), and a linear regression (R)



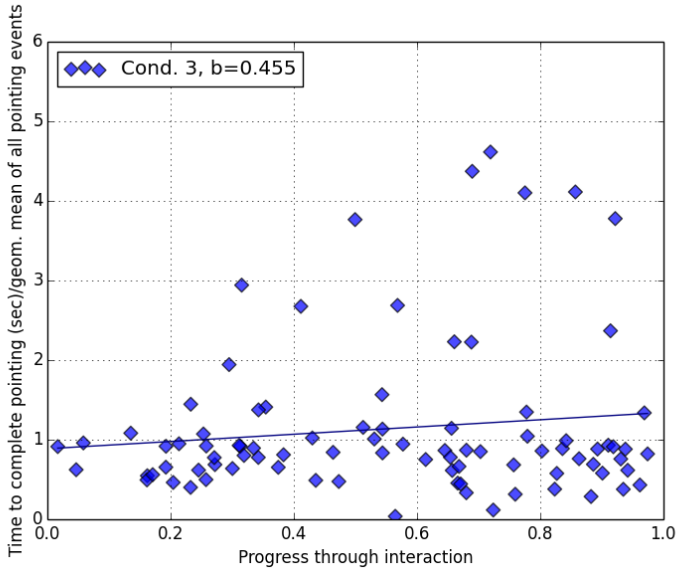
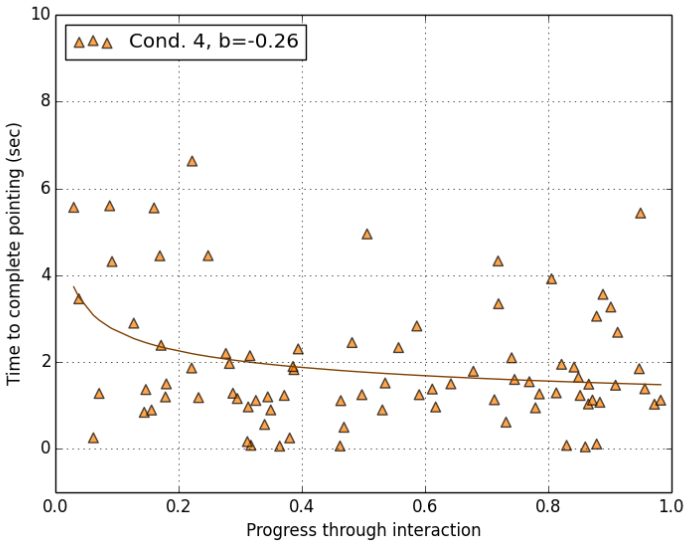




Figure: Results with table, hint given. $b_\rho \approx 0.315$, $s \approx 1.245$; $b_\mu \approx 0.455$

# Results

- X: % progress through trial; Y (L): time to complete successful pointing; Y (R): time to complete successful pointing, as ratio to the geometric mean of all user's recorded pointing times
- Best fit line is shown as a least-squares fitted power law (L), and a linear regression (R)



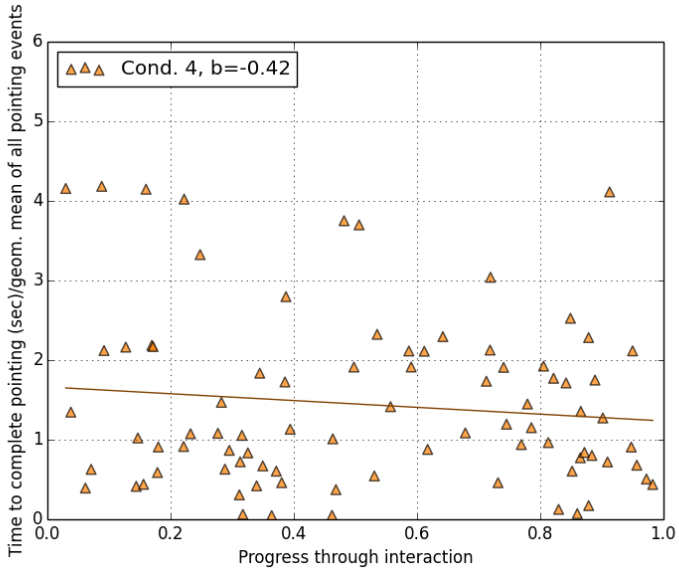




Figure: Results without table, hint given. $b_\rho \approx$ -0.265, $s \approx 0.832$; $b_\mu \approx$ -0.427

# Discussion

- Trend of increasing difficulty in successfully pointing in conditions with the table
- Trend of more efficient pointing in conditions without the table
- Opposite of what we expected
  - Table did not seem to provide the users with a reference point with which to ground deictic gestures
  - Seemed to make pointing more difficult
    - 1) Introduced a measure of confusion to the interaction
    - 2) Caused users uncertainty about valid reference points (i.e., table vs. screen)

# Discussion

- Where subjects were given more information (or hints), difference between "table" and "tableless" condition is more pronounced
- Nearly flat lines in Conditions 1 and 2 suggest users barely changed their pointing strategies at all
- Speculation: Users settle on a particular strategy (likely pointing at the screen/toward the avatar), and persist
- Subjects in Conditions 3 and 4, given hints about the table, display either marked adaptation (4) or marked confusion (3).
- Speculation: When attention was drawn to physical table, users tried to use it, got confused if they did not succeed at first
- Speculation: Without the table, users could more easily use the empty space to mirror coordinates in virtual world

# Discussion

- Table served as distractor
- Imposed extra cognitive load on task of trying to integrate real world with virtual world
- Reflects known difficulties in "mixed-reality" environments (Benford et al., 1998; Flintham et al., 2003)
  - Due to cognitive load of in transforming one's embodied coordinate system to virtual world
  - Further research needed into influence of exact instruction phrasing

# Discussion

- Other hypotheses:
  - (Sometimes) pointing became more difficult in later stages of the trial
    - As structure emerged, more precision required → more difficulty
    - May be overridden by adaptation in other conditions
  - Subjects allowed free reign to adapt overall strategy for building task
    - i.e., for actions supervenient on gestures such as pointing
    - Where pointing proved difficult, user might adapt by relocating items (by pointing), or loosening constraints on desired actions
    - e.g., allowing spaces between the blocks so that pointing at block locations would be easier

# Discussion

- Providing instructions led to more marked results than providing no guidance
- Suggests that the user's model of the situation matters, as well as the physical situation itself
- Small sample size due to partition ($N$=5)
- Tentative results
  - Evidence for switching implementation from table pointing to screen (complete)
  - Intriguing results may become more pronounced in studies with more subjects
- Deixis is just one part of interacting with a virtual world
- But important!
- Insights into how to treat deixis in a virtual environment should be useful to developers seeking to build intelligent systems capable of interacting fluently with humans

# Discussion

- Contrary to expectations, real table interfered with ease of pointing
- Human watching virtual environment creates simulation of virtual world (user's mental simulation)
- *Get rid of the table, allow me to simulate what it represents in the physical world!*
- Provides insights into complexity of simulation itself independent of integration of physical reality

# Acknowledgments

- Brandeis University collaborators: Kyeongmin Rim, Tuan Do
- Colorado State University collaborators: Prof. Bruce Draper, Prof. Ross Beveridge, Pradyumna Narayana, Rahul Bangar, Dhruva Patil, Gururaj Mulay, and Jason Yu
- University of Florida collaborators: Prof. Jaime Ruiz, Isaac Wang, and Jesse Smith
- DARPA Communicating with Computers program

# Thank You!