On the Human Capacity for Physical and Analogical Inference

Author
Kubricht, James Robert

Publication Date
2018

Peer reviewed|Thesis/dissertation
On the Human Capacity for
Physical and Analogical Inference

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Psychology

by

James Robert Kubricht

2018
The field of intuitive physics has been reinvigorated in recent years, providing converging computational, behavioral, and neural evidence demonstrating that humans possess a wealth of knowledge about the attributes and dynamics of physical entities in the world. However, progress in the field has led to many questions regarding (i) the human capacity for perception, physical inference, and higher-level reasoning; and (ii) how to design a machine to emulate newfound capabilities. My dissertation focuses on a range of research topics, from the characteristics of human perception and representation in dynamic scenes, to mental simulation of novel situations and transfer of higher-level knowledge to seemingly unrelated problem domains. In sum, my dissertation makes five major contributions:

1. There is a significant division between early research in intuitive physics and more recent investigations. This is largely due to a discrepancy between the reasoning systems utilized in each era’s tasks; namely, explicit reasoning in early (pencil-and-paper) tasks, and implicit reasoning in recent (prediction and recognition) ones. While many earlier results have been explained by modern computational approaches, a number of findings remain unexplained. My dissertation provides a review of early intuitive physics research and places it in the context of modern computational approaches utilizing Bayesian statistics and machine learning. Critically, my dissertation outlines the shortcomings of each approach and the hurdles that must be overcome before an intelligent and generalizable intuitive physics architecture can be developed.
2. We examine how humans perceive and represent observable information in dynamic, two-body object collision situations. The hypothesis that humans perceive motion in relation to meaningful—sometimes moving—landmarks in space is tested under the noisy Newton framework for mass and causal inference. We term this the relative noisy Newton model, with the key distinction that the slow motion prior for motion perception is evaluated with respect to stationary points in the local environment.

3. We report human experiments demonstrating the viability of virtual reality (VR) technology in intuitive physics research. VR technology allows for the construction of novel environments whose characteristics differ from the real world in ways that preclude laboratory experimentation. One manipulation explored in my dissertation is the variation of gravitational acceleration in projectile motion situations. We show that the influence of unfamiliar gravity fields on human reasoning varies based on task demands. Shortcomings of current VR technologies are further discussed.

4. We introduce the intuitive substance engine (ISE) for explaining human reasoning about non-solid substance dynamics through mental simulation. This model is formed under the noisy Newton framework and provides converging evidence that people reason about complex physical situations by propagating noisy representations forwards in time using approximate, Newtonian principles. We explore whether people utilize coarse perceptual approximations of non-solid substance volumes when making their predictions; specifically, by representing substances as collections of rigid balls. We further propose specific task and stimulus characteristics which facilitate scene representation and subsequent mental simulation.

5. We explore the benefits of animated instructional materials in demonstrating higher-level concepts based on low-level physical interactions. We report a moderated mediation (path analysis) model exploring how individual differences in fluid intelligence impact comprehension of source information and subsequent transfer to novel target problems. The reported findings are the first to explore the interaction between fluid intelligence and spontaneous analogical transfer.
The dissertation of James Robert Kubricht is approved.

Philip Kellman
Keith Holyoak
Demetri Terzopoulos
Hongjing Lu, Committee Chair

University of California, Los Angeles
2018
To my parents and siblings . . .

With much love, thank you for seeing me through this.
1 Introduction ................................................................. 1
   1.1 Representation of Perceptual Variables in Intuitive Physics ........ 4
   1.2 The Role of Mental Simulation in Intuitive Physics .................. 6
   1.3 Transfer of Physical Knowledge to Novel Situations .................. 7

2 Intuitive Physics: Current Research and Controversies ................. 9
   2.1 Abstract ......................................................... 9
   2.2 Reinvigorating Intuitive Physics .................................. 9
   2.3 Apparent Misconceptions in Intuitive Physics ...................... 10
   2.4 Noisy Newton Framework ........................................ 16
   2.5 Probabilistic Simulation Approaches to Intuitive Physics .......... 17
   2.6 Intuitive Physics with Liquids and Other Substances ............... 19
   2.7 Inference of Physical Variables .................................. 20
   2.8 Learning About the Physical World ................................ 21
   2.9 Concluding Remarks and Future Prospects ......................... 23

3 Representing Perceptual Variables ..................................... 24
   3.1 Abstract ......................................................... 24
   3.2 Introduction .................................................... 24
      3.2.1 The Noisy Newton Framework ................................ 26
   3.3 Experiment 1: Mass Judgments .................................... 28
      3.3.1 Method ....................................................... 28
      3.3.2 Results .................................................... 30
   3.4 Experiment 2: Causal Ratings ..................................... 32
3.4.1 Method ........................................ 32
3.4.2 Results ........................................ 33
3.5 General Discussion ............................... 36

4 Intuitive Physics in Virtual Reality .................. 37
  4.1 Abstract ........................................ 37
  4.2 Introduction ..................................... 37
  4.3 Background and Related Work ...................... 39
    4.3.1 Intuitive Physics .......................... 39
    4.3.2 Mental Simulation .......................... 41
    4.3.3 Virtual Reality ............................ 43
  4.4 Experiment Overview ............................. 44
  4.5 Experiment 1: Direct Action ..................... 46
    4.5.1 Method .................................... 46
    4.5.2 Results .................................... 49
  4.6 Experiment 2: Speed Judgment ................... 52
    4.6.1 Method .................................... 53
    4.6.2 Results .................................... 54
  4.7 Experiment 3: Contact Location Prediction ....... 57
    4.7.1 Method .................................... 57
    4.7.2 Results .................................... 59
  4.8 Experiment 4: Flight Duration Estimation ........ 62
    4.8.1 Method .................................... 62
    4.8.2 Results .................................... 63
  4.9 Comparison Between Experiments .................. 64
5 Reasoning Through Mental Simulation

5.1 Abstract

5.2 Introduction

5.3 Intuitive Physics as Probabilistic Simulation

5.3.1 The Noisy Newton Framework

5.3.2 Intuitive Substance Engine (ISE)

5.3.3 Perceptual Noise and Uncertainty

5.4 Experiment 1: Reasoning about the Relative Pouring Angle of Liquid-Filled Containers

5.4.1 Method

5.4.2 Results

5.5 Experiment 2: Reasoning about Granular Materials

5.5.1 Method

5.5.2 Results

5.6 Experiment 3: Reasoning about Complex Interactions Between Sand and Rigid Obstacles

5.6.1 Method

5.6.2 Results

5.7 General Discussion

5.7.1 Are Precise Numerical Simulation Methods Needed to Explain Human Judgments?

5.7.2 Implementing Mental Simulation Uncertainty Into the MPM

5.7.3 Concluding Remarks
6 Transferring Physical Knowledge ........................................... 104

6.1 Abstract ........................................................................ 104

6.2 Introduction .................................................................... 105

6.2.1 Spontaneous Analogical Transfer ................................. 106

6.2.2 Visuospatial Displays and Analogical Transfer .............. 107

6.3 Experiment 1: Animated Source Materials Facilitate Transfer ........................................... 109

6.3.1 Method ....................................................................... 109

6.3.2 Results ....................................................................... 113

6.4 Experiment 2: Source Understanding Questions Inhibit Transfer ........................................... 119

6.4.1 Method ....................................................................... 120

6.4.2 Results ....................................................................... 121

6.5 General Discussion ............................................................ 125

7 Conclusion ........................................................................ 128

7.1 Developing a Physical Model of the World ............................ 131

7.2 Do Explicit Conceptions Influence Physical Intuitions? ........ 133

7.3 Physical Judgments in Virtual Reality ................................. 135

A Background and Outstanding Questions (Chapter 2) .................. 138

A.1 Erroneous Theories of Motion .......................................... 138

A.2 Early Research on Intuitive Physics ................................. 139

A.3 Object Collision Physics ............................................. 140

A.4 Trends ........................................................................ 141

A.5 Outstanding Questions .................................................. 141

B Model Details (Chapter 3) ..................................................... 143
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Examples of Intuitive Physics Problems</td>
<td>11</td>
</tr>
<tr>
<td>2.2</td>
<td>Three Computational Approaches to Mass Inference</td>
<td>14</td>
</tr>
<tr>
<td>3.1</td>
<td>Collision Event</td>
<td>26</td>
</tr>
<tr>
<td>3.2</td>
<td>Moving Background Display</td>
<td>29</td>
</tr>
<tr>
<td>3.3</td>
<td>Human Data and Model Predictions (Experiment 1)</td>
<td>31</td>
</tr>
<tr>
<td>3.4</td>
<td>Human Data and Model Predictions (Experiment 2)</td>
<td>34</td>
</tr>
<tr>
<td>4.1</td>
<td>Illustration of Experiments</td>
<td>41</td>
</tr>
<tr>
<td>4.2</td>
<td>Experiment Designs</td>
<td>47</td>
</tr>
<tr>
<td>4.3</td>
<td>Human Results (Experiment 1)</td>
<td>50</td>
</tr>
<tr>
<td>4.4</td>
<td>Regression Results (Experiment 1)</td>
<td>50</td>
</tr>
<tr>
<td>4.5</td>
<td>Human Results (Experiment 2)</td>
<td>55</td>
</tr>
<tr>
<td>4.6</td>
<td>Regression Results (Experiment 2)</td>
<td>56</td>
</tr>
<tr>
<td>4.7</td>
<td>Human and Model Results (Experiment 3)</td>
<td>61</td>
</tr>
<tr>
<td>5.1</td>
<td>Illustration of Water Tasks</td>
<td>72</td>
</tr>
<tr>
<td>5.2</td>
<td>ISE Modules</td>
<td>78</td>
</tr>
<tr>
<td>5.3</td>
<td>Flow Demonstration Video and Judgment Trial (Experiment 1)</td>
<td>84</td>
</tr>
<tr>
<td>5.4</td>
<td>Human Results (Experiment 1)</td>
<td>85</td>
</tr>
<tr>
<td>5.5</td>
<td>Model Results (Experiment 1)</td>
<td>89</td>
</tr>
<tr>
<td>5.6</td>
<td>Pouring Demonstration Video and Judgment Trial (Experiment 2)</td>
<td>90</td>
</tr>
<tr>
<td>5.7</td>
<td>Human and Model Results (Experiment 2)</td>
<td>92</td>
</tr>
<tr>
<td>5.8</td>
<td>Judgment Trial for Each Substance (Experiment 3)</td>
<td>96</td>
</tr>
<tr>
<td>5.9</td>
<td>Model and Human Basin Predictions (Experiment 3)</td>
<td>99</td>
</tr>
</tbody>
</table>
6.1 Illustration of Scenario Diagrams ........................................ 111
6.2 Spontaneous and Total Transfer Rates (Experiment 1) ............ 114
6.3 Source-Understanding Scores (Experiment 1) ........................ 117
6.4 Source Understanding Scores Across RPM Group (Experiment 1) 118
6.5 Spontaneous Transfer Rate Across RPM Group (Experiment 1) ... 119
6.6 Moderated Mediation Analysis (Experiments 1 and 2) .......... 123
LIST OF TABLES

4.1 Regression Results (Experiments 1 and 2). . . . . . . . . . . . . . . . . . . . . . 51
4.2 Model Results (Experiment 3). . . . . . . . . . . . . . . . . . . . . . . . . . . . 62
4.3 Human Results (Experiment 4). . . . . . . . . . . . . . . . . . . . . . . . . . . . 64
4.4 Human Results Comparison (Experiments 3 and 4). . . . . . . . . . . . . . . . . 66
5.1 Model Performance Measures (Experiment 2). . . . . . . . . . . . . . . . . . . . 94
6.1 Cannon Scenario Attributes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110
ACKNOWLEDGMENTS

I am deeply indebted to those who have supported me throughout my time at UCLA. To my parents, thank you so much for being there to provide perspective and motivation when things were overwhelming. I have also been fortunate enough to work with several remarkable researchers over the past five years. To Dr. Hongjing Lu, I would not be where I am today if it were not for your guidance and support. You are a role model on many levels, and I am deeply grateful for your effort in helping me to improve and succeed. Thank you for introducing me to the field of intuitive physics and providing me with countless (exciting) research opportunities throughout my time at UCLA. To Dr. Keith Holyoak, I cannot put into words how much I have enjoyed working with you these past years. You have challenged me to see things from new perspectives, which I look forward to applying in the years to come. You have helped me to develop a broad understanding of human perception and reasoning, and for that I am incredibly thankful. To Dr. Phil Kellman, thank you for introducing me to characteristics of human perception that have inspired within me a great deal of fascination and curiosity. To Dr. Patricia Cheng, your support and input have been invaluable to my research. Thank you for your kindness—it has meant a great deal to me. To Dr. Demetri Terzopoulos, your input and support have been invaluable to this dissertation. To Dr. Yixin Zhu, you are a computational wizard. I have thoroughly enjoyed working on our projects together, and I wish you the best in your career moving forwards. To Dr. Chenfanfu Jiang, our work would not exist without your knowledge and expertise in the field of computer graphics. I am grateful for the opportunity to work with you on our substance dynamics projects. Also, to Dr. Gennady Erlikhman, you are an exceptional teacher and have contributed to countless lines of dialogue that have helped me to develop my ideas and research. Thank you sincerely for your contributions and support throughout my time at UCLA.

To my lab-mates, old and new: Drs. Junzhu Su, Jeff Bye, Derek Powell, and Melissa DeWolf; Yujia Peng, Junho Lee, Akila Kadambi, Maureen Gray, and Hannah Lee, thank you for giving me the opportunity to conduct research with such amazing people and for always
providing valuable input to my research during lab meetings. To my RAs: Cindy Xiong, Aditi Gala, Angela Cao, Tanner Velten, Marshall Briggs, Raphael Wang, Junling Dong, Sania Syed, it has been a pleasure performing research with all of you over the past five years. I would also like to thank my siblings for their unconditional support: Karah Varley, Chris Hinds, Corey Hinds, Connie Alexander, Kimberly Vargas, and George Alexander, you have always been there for me when I needed y’all, and I couldn’t ask for anything more. To Keela Thomson, you believing in me has helped me to keep moving forward—thank you sincerely for your unconditional motivation. To my friends, Cody Johnson, Carl Buist, Bo Braddock, Mac Xing, Jeff Chiang, Mark Straccia, Alex Siegel, Nick Baker, Suzy Carrigan, Rachel Older, Maxwell Mansolf, Alex Stolyarova, and Gene Ruby, you have always helped me to smile and remember to value the little things. It has been great having you along for the ride.

Portions of this work were supported by a NSF Graduate Research Fellowship, UCLA Graduate Division Dissertation Year Fellowship and Dean’s Scholar Award, NSF grant BCS-1353391, DoD CASIT grant W81XWH-15-1-0147, DARPA SIMPLEX grant N66001-15-C-4035, and ONR MURI grant N00014-16-1-2007.
VITA

2008–2012 Houston Endowment Scholar, Jesse H. Jones and Mary Gibbs Foundation

2011 National Undergraduate Research Fellow, NSF, PPPL, Dr. Hantao Ji

2012–2013 Research Assistant, UT Austin, Maddox Lab, Dr. Todd Maddox

2013 B.S. in Physics, UT Austin, Honors

2013–2015 Dean’s Scholar, UCLA Graduate Division

2013–2018 Graduate Research Assistant, UCLA, CVL Lab, Dr. Hongjing Lu

2013–2018 Graduate Research Assistant, UCLA, Reasoning Lab, Dr. Keith Holyoak

2014–2017 Graduate Research Fellow, NSF

2014 M.A. in Psychology, UCLA

2017 Ph.D. Candidate in Psychology, UCLA

2017–2018 Dissertation Year Fellow, UCLA Graduate Division

PUBLICATIONS * denotes joint first authors


xvi


CHAPTER 1

Introduction

People exhibit a remarkable capacity for making inferences about novel events and situations in the world. However, early research in the field of intuitive physics provided extensive evidence that people’s expectations do not always conform with normative (ground-truth) principles (e.g., Michotte, 1963; McCloskey, 1983; McCloskey, Washburn, & Felch, 1983; Todd & Warren Jr., 1982). These findings strongly conflicted with later developmental results demonstrating that humans attain a wealth of knowledge about how the physical world ‘works’ within the first year of life (Baillargeon, 1994; Kim & Spelke, 1999; Spelke, 1994). The picture that has emerged in the most recent decades is that people exhibit common misconceptions when reasoning about physical situations at the explicit level but can implicitly form accurate inferences when problems are presented in a natural, dynamic, and visually rich context (Smith, Battaglia, & Vul, 2013). Ironically, the static and impoverished (pencil-and-paper) stimuli utilized in several early intuitive physics tasks appear to have not been very intuitive at all.

Today, research has provided a breadth of evidence in support of the noisy Newton framework for physical inference, which posits that people reason about physical situations by combining noisy perceptual inputs with prior beliefs about latent physical attributes (e.g., object mass, substance viscosity, etc.) using approximated, ground-truth (Newtonian) constraints (Bates, Yildirim, Tenenbaum, & Battaglia, 2015; Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Sanborn, Mansinghka, & Griffiths, 2013). Although people’s expectations generally deviate from ground-truth outcomes, discrepancies are often explained by the influence of prior expectations on inferred perceptual and physical variables: e.g., the
slow and smooth prior for local motion signals (Weiss & Adelson, 1998). A key shortcoming of the noisy Newton framework, however, is that the governing laws of physics are explicitly written into computational models within task-specific domains. Consequently, models for one task typically cannot generalize to novel tasks involving physical entities whose mechanics are separately defined. Moreover, it is implied that people possess (and efficiently apply) mental models which approximately capture the underlying physical rules for combining spatially represented variables in a given situation. Although this appears reasonable for simple and/or common physical events—such as two-body collisions where the governing equations are linear and computationally tractable—it becomes less compelling when more complicated situations are considered (e.g., events involving non-solid substances; see Bates et al., 2015; Hespos, Ferry, Anderson, Hollenbeck, & Rips, 2016; Kawabe, Maruya, Fleming, & Nishida, 2015).

It remains unclear (i) how people construct mental models which emulate the principles of Newtonian physics, especially in the domain of substance dynamics; (ii) the degree to which those models capture physical complexity in the world; (iii) how different models are utilized in various types of intuitive physics tasks; and (iv) how people know to transfer and apply a given model in a novel situation. Moreover, there remains ambiguity in how perceptual inputs are represented prior to being passed to the reasoning system. For example, are the positions of objects perceived relative to a fixed point in the environment or in relation to moving entities which share the object’s motion (Clarke, Öğmen, & Herzog, 2016; Duncker, 1937; Kaiser, Proffitt, & McCloskey, 1985; Proffitt, Cutting, & Stier, 1979)? To address these questions, my dissertation examines the human capacity for inferring properties of novel dynamic events involving rigid objects, non-solid substances, and electromagnetic waveforms (i.e., radiation). The key questions addressed in the following chapters stem from the perception and reasoning components of intuitive physics. Regarding perception, what is the role of dynamic sensory information in the extraction of meaningful knowledge about the environment, and to what degree does the perceptual system approximate situations spatially? Then, how do those inputs interact with the reasoning system to form predictions and judgments about novel events? Chapters 3, 4, 5, and 6 are each relevant to the first
question. Chapter 3 examines how inferences about hidden physical attributes are formed from visual estimates of velocity in dynamic 2D displays, and Chapter 4 investigates how haptic information from virtual interaction within a 3D environment—or the lack thereof—guides motor planning and execution. Chapter 5 studies how estimates of non-solid substance attributes are attained from visual motion cues in realistic, dynamic imagery, and Chapter 6 examines how animated demonstrations of higher-level concepts can be leveraged to facilitate understanding and subsequent transfer to new situations. Regarding the second question, Chapter 5 provides a rigorous exploration of spatial and physical approximation in substance-related situations and is further elaborated upon in the following paragraph. Finally, the third question is addressed in Chapters 3, 4, 5, and 6. Chapter 3 examines how biased inferences about motion impact physical reasoning in object collision situations; Chapter 4 shows how biased human predictions, judgments, and interactions are explained by inferred estimates of critical hidden variables based on prior expectations and the absence of haptic experience; Chapter 5 also explains biases in human performance as resulting from biased physical inferences based on observable motion cues; and Chapter 6 shows how low-level visual information guides analogical reasoning.

Taken together, my dissertation aims to address the aforementioned topics from a behavioral and computational perspective. The following sections provide motivation for the experiments reported in the chapters herein, and a review of the field of intuitive physics is provided in Chapter 2. The review places recent literature in the context of classical results, outlining the current state of the field and outstanding questions moving into the future. Each of the following sections addresses a subset of the questions raised in the previous paragraph. Section 1.1 discusses how motion is represented in dynamic scenes: e.g., are moving objects represented with respect to stationary positions in the environment, or with respect to moving entities which share their motion? Chapter 3 further explores this question in the domain of two-body object collisions, testing the hypothesis that relative motion signals are utilized by the reasoning system when inferring mass and causality. Another important research question involves the spatial resolution or fidelity with which situations are represented—and whether coarse spatial and physical approximation can explain human
reasoning through mental simulation. Coarse approximation of non-solid substances and its role in mental simulation is further discussed in Section 1.2, and the human capacity for predicting the dynamics of non-solid substances through mental simulation is rigorously explored in Chapter 5. To mentally simulate dynamic physical events, one must possess accurate estimates of the attributes or characteristics of relevant entities. Dynamic visual information (e.g., viewing an event unfolding across time) is incredibly effective in providing this type of information. Chapter 4 demonstrates the viability of virtual reality (VR) technology in examining human perception and reasoning in artificial scenes following observation of dynamic events. However, dynamic information can also be leveraged to demonstrate higher-level concepts (or schemas) which can transfer to seemingly unrelated problems in dissimilar domains. Section 1.3 outlines how perception and physical reasoning tie into symbolic understanding, and Chapter 6 examines the human capacity for transferring higher-level principles—learned from animated instructional materials—to novel target problems. The corresponding experiments are the first to study the impact of individual differences in fluid intelligence on spontaneous analogical transfer. Finally, Chapter 7 summarizes the findings of the preceding chapters and proposes directions for future work.

1.1 Representation of Perceptual Variables in Intuitive Physics

Although many earlier findings demonstrating common physical misconceptions appear to have been due to impoverished task materials—such as static line drawings of an event at a single point in time—their results sometimes extended to experiments utilizing realistic, dynamic stimuli. One of the earliest examples is Piaget and Inhelder’s (1956) water-level task (WLT), in which children are shown a rotated, two-dimensional container printed on paper and are asked to draw the surface of the water inside. Interestingly, both children and adults draw water lines that deviate approximately 5° from the horizontal in the direction of the container’s rotation (Hecht & Proffitt, 1995; Howard, 1978; McAfee & Proffitt, 1991; Vasta & Liben, 1996; see Figure 5.1a). This finding extends from static (pencil-and-paper) to dynamic (motion display) tasks (Howard, 1978; McAfee & Proffitt, 1991), indicating an
underlying perceptual effect. McAfee and Proffitt (McAfee & Proffitt, 1991) proposed that this effect is due to representation of container contents based on *rotated* coordinate axes aligned with the container’s edges: i.e., one axis points in the cylinder’s radial direction, and the other axis lies parallel with the cylinder’s base. The frame of reference proposal is further reinforced by the finding that people who work in professions where containers are interacted with regularly are more susceptible to biased water-level estimates (Hecht & Proffitt, 1995).

Researchers have also examined the impact of *moving* frames of reference on people’s physical expectations. An important early example is the Gestalt demonstration of wheel motion put forth by Karl Duncker (Duncker, 1937). In the Duncker wheel animation, two discs are shown. One of the discs—perceived as the center of a wheel—moves along a linear path left to right, and the other disc—perceived as fixed to the outer rim of a wheel—moves along a cycloid path\(^1\). Rather than perceiving the cycloid path, people perceive the outer disc as rotating in a circle around the center disc. Later experiments suggest that relative motion (i.e., motion of the outer disc relative to the moving inner disc) is extracted early in the motion perception process (Proffitt et al., 1979). More recent computational approaches have provided explanations for how non-retinotopic reference frames are utilized when synthesizing motion in dynamic scenes (e.g., Clarke et al., 2016).

Another important example of relative motion perception involves the path an object will take towards the ground after it is dropped from a moving body (Howe, Tavares, & Devine, 2012; Kaiser, Proffitt, & McCloskey, 1985; Krist, 2000; McCloskey et al., 1983). The erroneous belief that an object will fall straight downward after it is dropped is encountered in both static and dynamic tasks (Howard, 1978; McCloskey et al., 1983). Although objects follow parabolic trajectories when dropped from moving bodies (relative to a stationary point in the environment), their motion relative to the moving body from which they are dropped is, in fact, straight downward. However, straight-down trajectory predictions were taken as evidence that people reason about projectile motion using an erroneous model of

\(^{1}\)An animation of the Duncker wheel demonstration can be viewed at the following URL: https://www.youtube.com/watch?v=AePyHt3Bblg
physics (see Appendix A). However, as the Duncker wheel shows, motion perception heavily relies on meaningful reference points in moving displays. Thus, when the moving body is removed from the falling-object situation, the straight down bias is alleviated (Kaiser, Proffitt, & McCloskey, 1985). This finding is crucial because it further demonstrates that people represent perceptual variables relative to meaningful reference frames that may vary from situation to situation. When modeling human perception and reasoning, it is vital to understand how perceptual input variables are represented before passing them to subsequent model components. A central aim of this dissertation is to test the hypothesis that human judgments about physical attributes and causality depend on relative motion information. This hypothesis is tested under the noisy Newton framework in the domain of two-body object collisions (Sanborn, 2014; Sanborn et al., 2013) with relative motion introduced via moving background referents.

1.2 The Role of Mental Simulation in Intuitive Physics

Representation of physical entities certainly extends beyond the reference frame utilized to capture their positions and movements across time. For example, people can infer the physical attributes of rigid objects (e.g., their mass) through visual texture cues (Gallivan, Cant, Goodale, & Flanagan, 2014), and non-solid substance attributes (e.g., viscosity) can be inferred from motion cues (Kawabe et al., 2015) and even static shape information alone (Paulun, Kawabe, Nishida, & Fleming, 2015). However, to infer the future status of a dynamic physical event, combination rules for position and attribute variable estimates are needed. This is problematic for non-solid substances because the combination rules for their dynamics are highly complex and operate on a large set of discretized substance elements (or material points; Jiang, Schroeder, Selle, Teran, & Stomakhin, 2015; Sulsky, Zhou, & Schreyer, 1995). On one hand, it is doubtful that the human perceptual system is capable of approximating non-solid entities as a collection of thousands of constituent elements. On the other hand, the partial differential (Navier-Stokes) equations governing substance dynamics are prohibitively demanding for computational systems, suggesting that people build
approximated mental models to simulate events forwards in time.

In my dissertation, I will explore the hypothesis that people reason about substance dynamics by first representing entities as a collection of rigid objects and then simulating their dynamics using basic Newtonian combination rules that coarsely approximate underlying viscous forces. Following previous findings indicating the human capacity for inferring substance dynamics (Bates et al., 2015; Hespos et al., 2016; Schwartz & Black, 1999), the work herein explores people’s expectations about viscous liquids—as well as granular materials—in three tasks adapted partially from previous research. In each experiment, candidate simulation models formed under the noisy Newton framework are compared with data-driven models which are naïve to underlying physical constraints. A primary aim of this research is to decipher whether human reasoning about substance dynamics can be captured by a coarse simulation model whose element-by-element interaction rules lend themselves to neural emulation (Battaglia, Pascanu, Lai, & Rezende, 2016; Chang, Ullman, Torralba, & Tenenbaum, 2016; Grzeszczuk, Terzopoulos, & Hinton, 1998).

1.3 Transfer of Physical Knowledge to Novel Situations

Intuitive physics is generally thought of as three sequential components: perception, followed by reasoning, and then interaction. The notion that the reasoning process operates as a black box equipped with a coincidentally relevant model of physics, however, is misleading. Instead, perceptual cues for meaningful physical relations appear to prompt observers to retrieve previous experiences (grounded in perception) and apply their corresponding physical models to form expectations about current situations (White, 2009, 2012). However, perceived surface similarity between previous (source) and present (target) situations is not necessarily a precursor to transfer. For example, interacting with a dynamic physical system in a goal-oriented manner can facilitate performance on a structurally analogous, non-physical problem (Day & Goldstone, 2011). This spontaneous noticing of relevant source experiences appears to occur at the implicit level, evidenced by relative worse performance in Day and Goldstone’s (2011) task when source-target mappings were explicitly considered. Then what
is the driving force behind the spontaneous, implicit component of analogical transfer? Is it the depth with which the source is understood, or are specific individual differences at play?

To answer these questions, my dissertation presents a statistical analysis of spontaneous analogical transfer rates to Duncker’s (1945) temporally dynamic radiation problem following observation of animated source instruction materials. Several earlier studies have explored spontaneous transfer to the radiation problem, generally using a semantically distant but structurally similar source analog (e.g., the military problem; Gick, 1985; Gick & Holyoak, 1980, 1983). Pedone, Hummel, and Holyoak (2001) later showed that animated displays are particularly effective in demonstrating the underlying convergence principle (i.e., multiple weak forces sum to a large force when they converge upon one another), arguably due to the principle’s temporally dynamic structure. A final aim of my dissertation is to systematically examine the impact of animated instruction materials on source understanding and subsequent analogical transfer, and to determine whether the impact stems from individual differences in performance on a fluid intelligence measure.
CHAPTER 2

Intuitive Physics: Current Research and Controversies

2.1 Abstract

Early research in the field of intuitive physics provided extensive evidence that humans succumb to common misconceptions and biases when predicting, judging, and explaining activity in the physical world. Recent work has demonstrated that, across a diverse range of situations, some biases can be explained by the application of normative physical principles to noisy perceptual inputs. However, it remains unclear how knowledge of physical principles is learned, represented, and applied to novel situations. In this review we discuss theoretical advances from heuristic models to knowledge-based, probabilistic simulation models, as well as recent deep-learning models. We also consider how recent work may be reconciled with earlier findings that favored heuristic models.

2.2 Reinvigorating Intuitive Physics

Humans are able to understand their physical environment and interact with objects and substances that undergo dynamic state changes, making at least approximate predictions about how observed events will unfold (e.g., predicting the trajectory of a thrown ball, the direction that a chopped tree will fall, or the path of a breaking wave). The knowledge underlying such activities is termed intuitive physics. This topic, an active research area for several decades, has recently been reinvigorated by new theoretical approaches linked to artificial intelligence. These theories have been used to model findings from behavioral studies that apply psychophysical measures to perception and reasoning with complex dynamic
displays. Here, we review recent research and theories (placing them in the context of earlier work), and discuss some of the ongoing controversies in the field.

2.3 Apparent Misconceptions in Intuitive Physics

Before the most recent decade, research on intuitive physics primarily focused on misconceptions that people demonstrate when reasoning about the attributes and movements of objects and substances in the world (e.g., McCloskey, Caramazza, & Green, 1980; McCloskey et al., 1983). Numerous studies found that humans exhibit striking deviations from Newtonian physical principles when asked to explicitly reason about the expected continuation of a dynamic event based on a static image representing the situation at a single time point. The predictions that people made in these studies often appeared to agree with erroneous theories of motion, rather than with (ground-truth) Newtonian physics (see Appendix A, Erroneous Theories of Motion). For instance, adults often predict that an object dropped from a moving body will follow a linear path downwards (i.e., the straight-down belief; McCloskey et al., 1983), and children predict that a horizontal force will propel a vertically moving object in the direction that it is pushed (DiSessa, 1982). Such evidence has been used to argue that people sometimes reason about the physical world using an Aristotelian model of physics. In other situations, adults appear to exhibit medieval impetus beliefs, for example, that an object exiting a curved tube will follow a curvilinear trajectory in the absence of external forces (McCloskey et al., 1980). In these studies (Figure 2.1; see Appendix A, Early Research on Intuitive Physics), participants were typically shown physical situations at a specific point in time via static diagrams printed on paper, and were required to draw how the situation would unfold going forwards in time. In such pencil-and-paper tasks, people often succumb to systematic errors when predicting the physical behavior of situations. People are not, however, internally consistent in their explicit reasoning judgments. Instead, they appear to reason in accord with different theories of motion depending on the situation (Cook & Breedin, 1994). These findings led to a generally pessimistic assessment of the human capacity to perceive and reason about physical situations, most notably in projectile motion and object collision situations (Todd & Warren Jr., 1982; Gilden & Proffitt, 1994).
Figure 2.1: Examples of Intuitive Physics Problems. The task in each problem is to reason about the attributes or movements of objects and substances in various situations. Aside from object collision judgments (a), problems have generally been depicted via a static diagram of the physical system. In (bd) the unbroken line corresponds to the correct trajectory, and the broken lines correspond to common, erroneous predictions. The probabilistic simulation framework has achieved success predicting people’s expectations about the attributes (a) and movements (c) of objects in dynamic displays, as well as the pouring angle of two fluid-filled containers (f). A computational account of people’s explicit trajectory predictions in (bd) has not been developed.

However, several studies have shown that these misconceptions can be reduced or even abolished when experimental paradigms and tasks are varied in particular ways (Kaiser, Jonides, & Alexander, 1986; Kaiser, Proffitt, & Anderson, 1985; Kaiser, Proffitt, Whelan,
Human errors are greatly reduced when explicit reasoning problems are presented in a familiar context (e.g., when an object exiting a curved tube is replaced by water exiting a curved hose; Kaiser et al., 1986). This finding suggests that specific prior knowledge can override more general physical intuitions that generate misconceptions. In addition, recent work has demonstrated that, although adults perform poorly when drawing the trajectory an object will follow after being released from a pendulum, they can successfully predict its landing location (Smith et al., 2013). It thus appears that systematic misconceptions are less likely to be exhibited in tasks that evoke implicit or tacit knowledge, such as recognizing the normative unfolding of an event in an animated display. In addition, the format of the stimulus display also influences how susceptible people are to misconceptions about object movements. For example, human judgments are more consistent with Newtonian physics when situations are presented in an animated format (Kaiser, Proffitt, & Anderson, 1985; Kaiser et al., 1992). Developmental studies have also yielded converging evidence that the judgments of children are more Newtonian when events are presented as animations (Kim & Spelke, 1999).

In addition to studies involving motion of a single object, early research about collisions between two objects also demonstrated deviations of human judgments from what is expected given Newtonian principles (see Appendix A, Object Collision Physics). These studies were inspired by classic work on perceptual causality which found that people report causal impressions that are not fully determined by the physical properties of the situation (Michotte, 1963; Sanborn et al., 2013). For example, consider the case in which an initially moving object (motor object) collides with one that is initially stationary (projectile object). When the physical effect of the motor object on the projectile object is relatively small (e.g., the post-collision velocity of the projectile object is less than the pre-collision velocity of the motor object), people report a stronger causal impression than when the physical effect is large (e.g., the post-collision velocity of the projectile object is greater than the pre-collision velocity of the motor object; Sanborn et al., 2013). In addition, when two equally heavy objects collide, people often report that the motor object weighs more than the projectile object, a phenomenon termed the motor object bias (Todd & Warren Jr., 1982). This find-
ing was traditionally interpreted in terms of heuristics: people may infer the attributes of colliding objects using two rules based on salient perceptual cues: (1) The object that moves fastest following a collision event is lighter (the velocity heuristic; Figure 2.2a); and (2) the object that deflects at the greatest angle is lighter (the angle heuristic; Gilden & Proffitt, 1994).

However, although these heuristics account for human judgments about the relative mass of colliding objects in some cases, they do not generalize to other situations. For example, it is unclear whether the heuristic model accounts for relative mass judgments when the motion of each object before the collision is occluded (Sanborn et al., 2013). The heuristic account also fails to explain why people are less susceptible to the motor object bias after completing a large number of training trials. Even if naïve observers use heuristic reasoning to make perceptual judgments, with experience they may transition to correct application of normative physical principles (Runeson, Juslin, & Olsson, 2000).

The general picture that emerges from research on intuitive physics is that people exhibit misconceptions and biases when (i) they are asked to provide explicit predictions or explanations about continuations of physical events, (ii) the events are unfamiliar and presented with minimal context, or (iii) the events are portrayed using impoverished stimuli, such as static line drawings depicting a situation at a single moment in time. Although misconceptions have often been attributed to people holding erroneous Aristotelian or impetus theories, it may well be the case that under these unfavorable conditions people do not employ a systematic theory to infer physical motion (Cook & Breedin, 1994). Instead, explicit predictions may draw upon a set of individualized background knowledge (Smith et al., 2013) based on salient perceptual cues that seem to be potentially relevant. Importantly, when these unfavorable conditions are alleviated (i.e., when people make more tacit judgments about familiar types of events based on rich visual displays), human judgments align more closely with Newtonian physics (Kaiser et al., 1986).

It thus seems possible that people in fact have a strong intuitive ‘physics engine’ available to them, but it is only evoked under favorable conditions. This does not imply, however, that people do not hold explicit conceptions of physical processes, nor that those conceptions
Figure 2.2: Three Computational Approaches to Mass Inference. (a) In the heuristic model, it is assumed that observed velocities are equivalent to physical velocities in the world. Post-collision velocities are compared, and the object that moves fastest following collision is assumed to be lighter. (b) The probabilistic simulation model places priors on physical variables. The motion prior biases perceived velocities towards slow motion. The likelihoods of different mass ratios are determined by comparing simulated final velocities to observed velocities. (c) In a deep-learning model, a convolutional neural network (CNN) is trained on 2D image inputs and outputs object attributes (mass and friction). The CNN is then used to predict object attributes from previously unseen image data.
cannot interfere with human reasoning in familiar environments. Indeed, people sometimes push objects along curved paths to impart upon them a curvilinear impetus (McCloskey, 1983), and even erroneously describe their own actions as adhering to the straight-down belief (McCloskey et al., 1983). However, cortical activation associated with explicit physical knowledge (Mason & Just, 2016) does not entirely overlap with brain activities associated with tacit physical inference (Fischer, Mikhael, Tenenbaum, & Kanwisher, 2016). Although people might plan their movements or describe their actions in accord with explicit physical conceptions, it appears that their tacit judgments and predictions are generally consistent with normative physical principles (Smith et al., 2013; Kaiser, Proffitt, & Anderson, 1985).

It remains uncertain how explicit physical conceptions are derived from experience and to what degree they interact with tacit physical knowledge (Howe, Tavares, & Devine, 2014). One reason for this uncertainty is the difficulty in classifying conceptions as arising from perceptual ambiguity or from ineffective representations of the perceptual and physical variables involved in a task. For example, the straight-down belief may appear to arise when drawing the trajectory of an object after being released from a pendulum because the speed of the object at the indicated (static) location is ambiguous. Alternatively, people might believe that an object dropped from a moving body will fall straight downwards because they represent the perceived motion of the object relative to the moving body. The finding that humans are less susceptible to the straight-down belief when the moving body is removed from the situation (Kaiser, Proffitt, & McCloskey, 1985) is consistent with the latter possibility. People are also commonly biased when drawing the water level on a rotated container (i.e., the water-level task; Piaget & Inhelder, 1956), even after explicitly stating that the surface of a liquid should remain horizontal regardless of the orientation of its container (Howard, 1978). In this case, the bias appears to arise due to ineffective representation: specifically, using an object-centered reference frame with axes parallel to the surfaces of the container (McAfee & Proffitt, 1991). Although such biases might be interpreted as predictions based on erroneous physical theories, they might instead reflect correct application of normative principles to variables represented in moving or rotated frames of reference. It is also possible that meaningful physical quantities (e.g., the mass distribution of a rotating wheel) are simply not
represented or are coarsely approximated in some situations, leading to erroneous predictions and judgments (Proffitt, Kaiser, & Whelan, 1990). People’s judgments about physical quantities (e.g., the forces that two colliding objects impart upon one another) have also been shown to disagree with fundamental Newtonian principles (White, 2009). Thus, it is important that future work in intuitive physics considers (i) correspondence between cognitive and physical constructs, (ii) the nature of cognitive representations across dissimilar problem contexts, (iii) the role of physical approximation in complex displays, and (iv) the interaction between explicit conceptions and tacit understanding in prediction and judgment tasks.

2.4 Noisy Newton Framework

In recent years, research on intuitive physics has been reinvigorated by new theoretical approaches based on Bayesian inference, most notably the noisy Newton framework, which integrates ground-truth physical principles with uncertainty about sensory information (Sanborn, 2014; Sanborn et al., 2013). Models based on the noisy Newton framework assume that people integrate noisy sensory inputs with prior beliefs about perceptual and physical variables underlying physical situations, and model the constraints among those variables in accord with Newtonian physics. In the case of collision events, predictions are modeled by simulating thousands of physical situations. In each simulation, the physical outcome is computed using Newtonian laws operating on sampled variables of perceptual and physical properties. Although most perceptual variables appear to be observable (e.g., velocity, location), it remains necessary to convert objective evidence (observations) into subjective estimates by integrating noisy sensory input with priors on statistical regularities of perceptual cues in the world. Another source of sampling uncertainty comes from unobservable physical properties (e.g., mass, viscosity) which must be inferred from sensory observations and/or general knowledge about the physical world (Figure 2.2b). The noisy Newton framework effectively reconciles several inconsistencies between human judgments and Newtonian physics (Sanborn et al., 2013).
2.5 Probabilistic Simulation Approaches to Intuitive Physics

Motivated by the initial successes of the framework, several researchers have recently extended the noisy Newton approach to explain human judgments and predictions about a variety of physical situations. As an overarching framework, the Bayesian approach to intuitive physics can be viewed as a tool for understanding how abstract knowledge—represented by priors and a generative function to assess likelihood—guides inference about object states from incomplete and noisy information about perceptual and physical variables (Tenenbaum, Kemp, Griffiths, & Goodman, 2011). In the noisy Newton framework, inference is achieved by passing noisy information through a physics engine, which is defined by the principle of conservation of momentum in the case of object collisions (Sanborn et al., 2013). Knowledge about object dynamics is ‘written into’ the model under the assumption that the transformation of perceptual inputs into physical expectations accords with Newtonian physical constraints.

The key idea underlying probabilistic simulation models is that humans construct mental models about physical situations, allowing for inference of future object states through mental simulation (Battaglia et al., 2013; Hamrick et al., 2016). The role of mental simulation is supported by work on mechanical reasoning which has demonstrated that people reason about physical systems by constructing and transforming spatial representations to answer questions about the behavior of objects and substances (Hegarty, 2004; Schwartz & Black, 1996, 1999). Spatial representation implies that object locations, motions, and hidden attributes in the physical world—as well as their interactions—are encoded and represented in the mind (Markman & Dietrich, 2000). Recent neural evidence suggests that the mental simulation process is likely carried out in cortical regions that overlap with the domain-general ‘multiple demand’ system of the brain (Fischer et al., 2016).

Probabilistic simulation models make judgments in physical reasoning tasks by integrating noisy information processing with advanced physics-based graphics engines to simulate future object states. In each simulation, the values of perceptual and physical variables in a scene are sampled according to distributions that emulate noisy information processing of
the positions, velocities, and attributes of the objects. Based on sampled states of perceptual and physical inputs, an ‘intuitive physics engine’ which approximates Newtonian principles is used to simulate future object states. The outcome of each simulation is then queried to form a predicted judgment, such as whether or not a tower of blocks fell down (Battaglia et al., 2013) or how much liquid fell into a designated area (Bates et al., 2015). Judgments are then aggregated across simulations to form a predicted response distribution. Parameters in the simulation model are chosen such that the distribution accurately reflects human behavior.

The probabilistic simulation approach has demonstrated promising results across several physical domains. Most studies utilizing probabilistic simulation examine the correlation between human performance and model predictions across a range of experimental conditions, rather than absolute performance levels (the typical focus in earlier studies). Overall, model predictions correlate well with human responses about the motions and attributes of stacked blocks (Battaglia et al., 2013; Hamrick et al., 2016) and about how liquids move past obstacles (Bates et al., 2015). Moreover, it has been shown that causal judgments about object collision outcomes are correlated with counterfactual assessments, specifically whether and how an outcome occurs (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Gerstenberg et al., 2017). The approach has also been used to explain human inferences about whether different containers can hold particular objects (Liang, Zhao, Zhu, & Zhu, 2015), as well as infant reasoning about complex displays of moving objects (Téglás et al., 2011). It is important to note that these simulation models account for human predictions primarily through the implementation of noisy information processing—the physical principles underlying object and substance dynamics are approximated but are not systematically biased.

However, uncertainty in the inferential mechanism (i.e., the physical model itself) may also constitute an important component in the intuitive physics ‘module’ of the brain. It has been shown that implementing randomness into the dynamics of objects in physical situations (on top of perceptual noise) provides a better fit to human performance in predicting the trajectory of an occluded object bouncing within a box (Smith & Vul, 2013). It appears that mental simulation outcomes are not deterministic for humans; instead, intermediate object
states are randomly perturbed during the inference process. Recent work suggests that, as the outcome of an event becomes increasingly uncertain or a problem becomes more difficult, additional cognitive resources are allocated to the mental simulation process (Hamrick, 2015).

2.6 Intuitive Physics with Liquids and Other Substances

Research on intuitive physics has transitioned beyond the behavior of solid objects to examine the human capacity to reason about the dynamics of liquids. This ability appears to emerge in the earliest stages of life because 5-month-old infants are able to distinguish between solids and non-solid substances (Hespos et al., 2016). A recent study showed that a computational model based on probabilistic simulation can account for human performance in predicting the resting configuration of a liquid pouring past obstacles into two empty basins (Bates et al., 2015). This model included perceptual noise applied to the initial locations of liquid elements, and simulated their movements using approximated normative physical principles. Probabilistic simulation was also employed to model performance in an explicit reasoning task concerning the angle at which a liquid-filled container will begin to spill through mental simulation (Kubricht et al., 2016). This paradigm is a variant of an earlier experiment in which participants reasoned about the angle at which two water-filled containers—one wider than the other—would begin to spill (Schwartz & Black, 1999; Figure 2.1f; see Appendix A, Early Research on Intuitive Physics). When asked explicitly to reason about which container would pour at a lesser angle, most participants mistakenly chose to rotate the wider container further. However, when asked to reason about the pouring angle of each container independently through imagined action, participants correctly rotated the narrower container further. In the modified task (Kubricht et al., 2016), two containers were filled with fluids varying in their viscosity (i.e., the apparent thickness or stickiness of a fluid), and visual motion cues based on flow visualization animations (Kawabe et al., 2015) were used to infer the viscosity of each fluid. Viscosity inferences were consistently biased towards lower values of viscosity, suggesting that people have a prior belief that fluids tend to behave like water. Results from the viscous fluid-pouring task indicated that human
judgments about the pouring angle of the two containers are consistent with a probabilistic simulation model utilizing normative physical principles given noisy perceptual inputs.

2.7 Inference of Physical Variables

The probabilistic simulation approach builds on two basic components: physical variables provided as the input to a physics engine, and physical principles encoded in the engine. Some physical variables (e.g., velocity and object positions) can be directly perceived, although perceived values could be distorted by neural noise and by generic priors (e.g., the slow and smooth prior in motion perception; Weiss & Adelson, 1998; Lu & Yuille, 2006). However, other physical variables (e.g., mass, viscosity, density, and gravity) are not directly perceivable. How could humans infer these physical attributes from low-level visual features of images?

Recent advances in deep learning models suggest a potential computational mechanism for inferring physical attributes from visual inputs and making predictions about physical situations. This approach arose in the field of machine learning, and is based on implementations of convolutional neural networks (CNNs; LeCun & Bengio, 1995; LeCun, Bengio, & Hinton, 2015; Krizhevsky, Sutskever, & Hinton, 2012). These networks take images encoded at the pixel level as inputs, and process the information through hierarchical layers to learn representations at multiple levels of abstraction, ranging from simple visual components (e.g., edges) to more complex patterns and object categories. A hybrid approach—integrating a knowledge-based physics model with a learning-based recognition network for predicting physical attributes from visual inputs—has had some success in accounting for human intuitive physical predictions (Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015). Utilizing a deep-learning network, dynamic visual inputs (sequences of 2D images) are mapped to inferred attributes (mass and friction) of two colliding objects through multiple processing layers (see Figure 2.2c). This procedure effectively inverts a key component of the generative physical process. The network is trained on image data tied to object attributes, which are determined by matching key features of the visual inputs to simulation output from a physics
engine. This model performs with an accuracy comparable to that of humans, demonstrating that learning-based methods can be effectively integrated with a knowledge-based physics engine to infer the attributes and dynamics of objects in the environment.

2.8 Learning About the Physical World

Approaches to physical reasoning based on probabilistic simulation typically assume that ground-truth physical principles are provided as prior knowledge. Thus, it is important to examine how such knowledge can be acquired. Is the mind of a child akin to a blank notebook with minimal information, as Alan Turing surmised (Turing, 1950), or is a cognitive architecture in place that guides the developing mind in learning about the physical properties of objects and the nature of their interactions?

Converging research on infants, children, and non-human primates has provided support for the core knowledge thesis (Spelke & Kinzler, 2007), according to which humans utilize separable systems of innate core knowledge, and these serve as building blocks for later learning. The thesis states that core physical principles guide the construction of tacit theories of motion (Baillargeon, 2002). Perceptual information serves as evidence for preexisting theories, and is represented with increasing complexity as theoretical understanding develops (Baillargeon, 1994). For example, infants first appear to grasp qualitatively whether a box and table are in contact with one another, and later become sensitive to the quantitative proportion of contact between the two surfaces. This variable identification process is repeated for different phenomena at various developmental stages, leading to piecemeal knowledge about physical situations which does not necessarily transfer to new situations that seem to be unrelated at the surface level (Baillargeon, 2002). In other words, initial knowledge about the physical world is specific to learned domains (Spelke, 1994).

From a computational perspective, one basic approach to learning is based on exemplars. Observed instances of a physical situation are represented as vectors in an N-dimensional space linked to corresponding attributes (e.g., whether the motor or projectile object in a collision is heavier; Cohen, 2006). Expected attributes of newly observed instances are predicted
by summing similarity measures across instances belonging to each possible classification. However, although the exemplar approach makes sensible predictions within constrained physical regimes by imitating physical knowledge, it fails to generalize to previously unseen regions in the stimulus space.

A more recent learning-based approach has been to directly emulate physical principles using deep-learning methods. The NeuroAnimator model (Grzeszczuk et al., 1998) is a neural network that emulates the mapping function which propagates a physical state forwards in time by viewing several instances of physical state transformations. Whereas the probabilistic simulation approach utilizes closed-form physical expressions to propagate scene states forwards in time, the NeuroAnimator model can achieve comparable performance by learning state transition patterns and applying them to previously unseen situations. When a general-purpose engine that emulates the laws of physics was trained on several physical domains, it then generalized to novel systems with different object quantities and relational rules (Battaglia et al., 2016). This is a promising step towards developing a learnable physics engine that can generalize to novel situations in a manner consistent with human abilities. Similarly, the PhysNet model, with an architecture based on a CNN, has achieved success in making physical predictions about simplistic block tower scenarios (Lerer, Gross, & Fergus, 2016). After training on artificial scenarios, PhysNet is capable of reasoning about the outcome and future trajectories of both artificial and real-world block tower configurations. It can generalize to new block-tower configurations with a different number of blocks than in the training cases, and yields predictions that are reliably correlated with human responses. However, the PhysNet model would have difficulty generalizing to situations that are more dissimilar to the training examples (e.g., the trajectory of a thrown object). In addition, unlike developing infants, the model requires several thousands of training examples to abstract basic physical knowledge about the environment. Future research will determine whether learning-based pattern recognition networks can be utilized to extract generalizable physical knowledge from perceptual inputs.
2.9 Concluding Remarks and Future Prospects

The field of intuitive physics in the past three decades has benefited from advances on several fronts: stimulus displays (from static diagrams to vivid dynamic animations controlled by computer graphics), computational theory (from heuristic accounts to a parsimonious framework based on probabilistic mental stimulation), and choice of physical situations to study (from a near-exclusive focus on the movement of solid objects to the behaviors of non-rigid fluids). The field now provides a model domain for quantitatively exploring the complex interrelationships between perception and reasoning.

Work guided by knowledge-based and learning-based approaches to intuitive physics suggests that a human intuitive-physics module integrates perceptual and reasoning processes to infer the behavior of physical situations. However, a great deal of future work will be necessary to develop a learnable and generalizable model of intuitive physical inference (see Appendix A, Outstanding Questions). Research on probabilistic simulation indicates that human predictions about physical situations are consistent with probabilistic inference. Nevertheless, such models require a vast number of simulations based on ‘hard-coded’ normative physical constraints to generate predicted response distributions (Bates et al., 2015; Battaglia et al., 2013; Gerstenberg et al., 2015; Hamrick et al., 2016; Sanborn, 2014; Sanborn et al., 2013; Smith et al., 2013). On the face of it, such computational complexity appears prohibitively demanding for the cognitive system (Davis & Marcus, 2015). Moreover, these models do not provide a full account of how physical principles can be learned through experience (Lake, Ullman, Tenenbaum, & Gershman, 2017), nor of how such principles might be implemented in neural circuitry (Fischer et al., 2016). Nonetheless, recent advances constitute progress towards developing a machine that can perceive, reason about, and interact with physical entities.
CHAPTER 3

Representing Perceptual Variables

3.1 Abstract

Human judgments about the physical attributes of—and causal relationship between—two colliding objects have been studied extensively over the past seventy years. Recent computational evidence suggests that judgments about the mass ratio of two colliding objects, as well as their perceived causal relation, can be explained by a coherent framework based on a Newtonian physical model and probabilistic inference resulting from noisy observations of object movements. However, it remains unclear how the physical and causal reasoning systems interact with the motion perception system when forming these judgments. The current study aims to examine whether high-level judgments are guided by object motion represented as relative motion with reference to a moving background, or as absolute motion with reference to a stationary position in the world. Both experimental evidence and model simulation results support the notion that physical and causal inference in object collisions depend on relative motion rather than absolute motion.

3.2 Introduction

Over the past seventy years, researchers have examined how human inferences about the attributes of—and causal relationship between—colliding objects vary according to spatiotemporal properties in observed displays (Cohen, 2006; Gilden & Proffitt, 1994; Leslie, 1982; Michotte, 1963; Natsoulas, 1961; Runeson, 1983; Runeson et al., 2000; Saxe & Carey, 2006; Schlottmann & Anderson, 1993; Scholl & Nakayama, 2002; Todd & Warren Jr., 1982;
White, 2006). In a typical launching event, an initially moving disc (motor object; or Object A) collides with an initially stationary one (projectile object; or Object B) and causes it to move forwards in the direction that it is pushed. Physically, the motor object interacts with the projectile object by imparting its momentum upon it; the sum of the motor and projectile objects’ momentum remains constant over time: i.e., the principle of conservation of momentum.

Michotte (1963) found that when people observe launching events, they report an immediate and irresistible impression that the motor object causes the projectile to move forwards. However, participants’ causal ratings were consistently attenuated when either (1) a spatial gap was placed between the colliding objects; (2) the projectile object’s movement was delayed; or (3) the projectile object moved faster/slower than the motor object following impact (see Figure 3.1). These findings indicate that causal impressions are highly sensitive to the spatiotemporal characteristics of observed events. However, Runeson (1983) later pointed out that causal impressions were too subjective to reliably measure and instead turned his attention towards relative mass judgments. Following Gibson’s (1966) doctrine of direct perception, he theorized that if people reason according to the principle of conservation of momentum, their judgments about which of two colliding objects is heavier should solely depend on their unbiased and accurate estimates of each object’s pre- and post-collision velocity.

Runeson’s predictions were subsequently tested by Todd and Warren (1982) who found that people are instead consistently biased towards reporting that the motor object is heavier (i.e., the motor object bias). Moreover, people are more susceptible to this bias when the objects are relatively inelastic: e.g., deformable vs. rigid balls. To explain these findings, researchers posited that observers form judgments according to simplified heuristic rules based on salient perceptual cues: e.g., which object moves faster after impact and the degree to which each object deflects off of the other (Gilden & Proffitt, 1994; Runeson et al., 2000). Although the heuristic approach qualitatively explains trends in the reported behavioral data, a more recent approach has demonstrated that people do appear to reason about relative mass in accordance with the principle of conservation of momentum, given that
Figure 3.1: **Collision Event.** Three common manipulations to spatiotemporal properties of a collision display in causal perception tasks. The dashed grey lines correspond with different time points during the collision event, and the arrows attached to the discs indicate their magnitude and direction of movement (i.e., velocity). (a) When a spatial gap is placed between the inside edges of two colliding discs at the moment of impact, perceived causality diminishes. (b) Introducing a pause (temporal delay) between when the motor object stops and the projectile object begins moving also attenuates perceived causality. (c) The numbers (1 or 2) indicate two possible outcomes; the projectile object moves either slower than the motor object following collision (Outcome 1) or it moves faster (Outcome 2). People report greater causal impressions after observing Outcome 1 relative to Outcome 2.

their perception is prone to error and prior beliefs about informative physical variables are held: i.e., the *noisy Newton* hypothesis (Sanborn et al., 2013; Sanborn, 2014).

### 3.2.1 The Noisy Newton Framework

The noisy Newton framework was proposed to explain people’s predictions about dynamic physical situations without the implementation of arbitrary heuristic rules. It has been employed across a wide range of physical domains, ranging from the movement of non-solid substances to causal reasoning through counterfactual simulation (see Kubricht, Holyoak, & Lu, 2017 for a review). The framework supposes that people possess an intuitive physics
“engine” (Battaglia et al., 2013) encoded in neural circuitry (Fischer et al., 2016) which approximately emulates the laws of physics to simulate spatially represented variables forwards in time (see Battaglia et al., 2016; Chang et al., 2016; Grzeszczuk et al., 1998 for a computational approach). Moreover, since people’s observations are inherently noisy, inferred estimates of observable variables are consistently biased towards prior expectations.

In the case of object collisions, Sanborn et al.’s (2013) implementation of the noisy Newton framework adopts the generic prior in motion perception (i.e., favoring slow motion in object motion) and a likelihood function which compares the observed velocity with the derived velocity from a physical model. The noisy Newton approach explains the motor object bias and predicts larger biases for relatively inelastic collisions. In addition, the noisy Newton model can be extended to account for causal judgments by comparing how well a noisy Newtonian model explains observations compared with a non-physical model (Sanborn et al., 2013; see Appendix B for model details).

Importantly, the noisy Newton framework does not specify how observable input variables should be represented: e.g., in a dynamic stimulus, object motion can be represented as relative motion with reference to a moving background, or as absolute motion with reference to a stationary position in the world. Furthermore, the physical inference may vary depending on whether the reasoning system adopts relative or absolute motion signals. For instance, imagine that you are looking out of the left window of a resting train and you see a vehicle move from your left periphery towards a second vehicle parked on the road nearby. The two vehicles collide, and the second vehicle correspondingly moves in the direction that it was pushed. You might get an impression that the two vehicles were equally heavy; but what if the train was traveling in the same direction as the initially moving vehicle? What if it was traveling in the opposite direction? Would your judgment about the weight of the two vehicles change? Would you be equally likely to report that the first vehicle had launched the second one forwards? Motion perception studies have shown that humans can perceive both relative and absolute motion with different degrees of sensitivity (Smeets & Brenner, 1994). For cognitive tasks probing the ability of physical and causal reasoning, it is important to understand what perceptual variables are selected and used for high-level judgments.
The central effort of the current experiments is to examine what motion information is
extracted from visual inputs for physical and causal judgments. Previous work on object
collision judgments have exclusively used stationary backgrounds, providing no distinction
between relative and absolute motion. However, in daily life, perceived landmarks are con-
stantly moving across our visual field as we move through—and interact with—the environ-
ment. In such cases, representing motion relative to those moving landmarks could provide
a different explanation of physical dynamics than absolute motion does. Across two exper-
iments, we (i) measured human performance in physical and causal judgment tasks when
viewing object collisions on a moving background, and (ii) compared noisy Newton model
predictions given absolute and relative motion inputs to test the hypothesis that humans
encode relative motion when forming mass judgments and inferring causality.

3.3 Experiment 1: Mass Judgments

The goal of the first experiment was to determine (1) whether a vertical background grid
moving with or against the motion of two colliding objects influences mass ratio judgments;
and (2) if so, whether the noisy Newton model for mass collisions with relative motion inputs
can explain participants’ performance.

3.3.1 Method

Participants. A total of 20 undergraduate students (14 female; Mean age = 21.2) were
recruited from the University of California, Los Angeles (UCLA) Department of Psychology
subject pool and were compensated with course credit.

Materials and Procedure. Collision event videos were presented on a 19” Dell E198WFP
LCD monitor with a refresh rate of 40 Hz at 1440 × 900 resolution. Videos were viewed
at a distance of approximately 70 cm. In each video, an initially moving object (termed as
motor object) collided with a stationary \( (u_B = 0) \) object (termed as projectile object). The
pre-collision velocity of the motor object varied across eight values: \( u_A = 1.9, 2.3, 2.6, 3.0, \)

28
3.4, 3.7, 4.1, and 4.5 cm/sec. The final velocities of the two objects were determined by Newtonian principles using a fixed restitution value of $e = 0.9$ and eight mass ratio values: $m_A/m_B = 1/3, 1/2, 2/3, 4/5, 5/4, 3/2, 2/1, 3/1$. In each video, the background also either moved against the direction of the collision (leftward; -2 cm/sec), with the direction of the collision (rightward; 2 cm/sec), or it remained at rest (0 cm/sec; see Figure 3.2). These manipulations yielded $8 \times 8 \times 3 = 192$ collision stimuli presented in a within-subjects design. Trials were presented in a randomized order, and no feedback was provided. Each stimulus video lasted 4 sec with impact occurring 2 sec into each collision event; the impact location was always at the center of the display. The collision videos were rendered using MATLAB Psychophysics Toolbox 3. The motor and projectile objects were depicted as black (RGB = 0 0 0) discs with 2.7 cm diameter. The vertical grid lines spanned the height of the screen (25.4 cm) and were colored gray (RGB = 150 150 150). Each line was 0.08 cm wide with a horizontal line separation of 2.7 cm.
Prior to the testing trials, participants were informed that they would be watching a series of videos where two discs interact with one another. They were told that there would be vertical lines behind the two discs in each display and that they would either move leftward/rightward or remain at rest. In each trial, participants viewed a collision video and then reported which of the two objects (left or right) they thought was heavier. Participants were provided with the opportunity to take two breaks which occurred 1/3 and 2/3 of the way through the experiment, which lasted approximately 20 minutes.

3.3.2 Results

Human Results. The proportion of participants choosing the motor object as appearing heavier in each mass ratio and background movement condition is displayed in the left panel of Figure 3.3. Participants’ responses—either 0 or 1—were averaged across the pre-collision motor velocity ($u_A$) conditions prior to analysis. A two-way repeated measures ANOVA was conducted on the response proportions to determine whether mass ratio and background movement influenced mass judgments. Results from the analysis indicated a significant interaction between mass ratio and background movement, $F(14,6) = 7.37, p = .01$, indicating that the impact of background movement on mass judgments varied according to mass ratio. As evident in Figure 3.3, participants were more likely to report that the motor object was heavier than the projectile object when the background moved leftward against the direction of the collision (Figure 3.2A), and less likely when it moved rightward in the same direction (Figure 3.2C). In other words, the point of subjective equality (PSE; i.e., the mass ratio where each judgment is equally likely) occurred at a minimum mass ratio with leftward background movement, PSE = 0.62, a moderate ratio when the background was at rest, PSE = 0.70, and a maximum ratio with rightward background movement, PSE = 0.99. In the following section, we explore whether the noisy Newton model for mass ratio judgments can explain this behavioral trend.

Model Results. The noisy Newton model for mass ratio judgments (Sanborn et al., 2013) takes as input the velocities of the motor and projectile objects and outputs the likelihood
Figure 3.3: **Human Data and Model Predictions.** The measured (left panel) and model predicted (right panel) proportions of participants choosing the motor object as appearing heavier. Separate lines indicate whether the background (BG) moved leftward/rightward or remained at rest. Proportions are averaged across pre-collision velocity ($u_A$) conditions.

of an observer choosing the motor object as appearing heavier. In the original noisy Newton model, the input of observed velocity is specified relative to a fixed point on the display; we will refer to these velocities as *absolute* velocities. This model can account for the motor object bias and predicts that the probability of choosing “motor object heavier” changes as a sigmoid function of the true mass ratio. The noisy Newton model with absolute velocity inputs is represented by the black curve in Figure 3.3 (right panel) and reveals a fit of $r^2(22) = 0.91$ (95% CI = [0.80, 0.95]). However, critically, the model’s performance is not influenced by the presence and direction of background movement—nor do the model predictions differ—since the absolute velocity inputs do not change across the three background movement conditions. Therefore, the model predicts the same PSE in each background condition, PSE = 0.95.

Alternatively, the noisy Newton model can take as input the motor and projectile velocities specified *relative* to a moving point fixed to the background grid. The result is relatively large velocities when the background moves leftward and relatively small velocities when it moves rightward. Since observation noise in the noisy Newton model increases for
larger velocity magnitudes, the influence of the slow motion prior is greatest in the leftward background condition and smallest in the rightward background condition. As shown by the separate curves in the right panel of Figure 3.3, the noisy Newton model with relative velocity inputs explains people’s increasing bias towards reporting “motor object heavier” in the leftward versus rightward background movement condition. The model also provides a superior fit to human judgments, \( r^2(22) = 0.97 \) (95% CI = [0.91, 0.98]), and predicts human PSEs: Leftward PSE = 0.54, Rest PSE = 0.95, Rightward PSE = 1.09. The model results for Experiment 1 used the same parameters reported in Sanborn et al. (2013): i.e., \( \sigma = 2 \), \( k_v = .1 \), \( w_v = .15 \).

### 3.4 Experiment 2: Causal Ratings

Our first experiment showed that the magnitude of the motor object bias depends on the background movement direction in a collision event. The noisy Newton model with relative motion inputs accounts for human mass ratio judgments well across a range of testing conditions. The purpose of the second experiment was to determine whether the same background manipulation affects perceived causality, and whether the noisy Newton model can account for human performance.

#### 3.4.1 Method

**Participants.** A total of 29 undergraduate students (20 female; Mean age = 20.5) were recruited from the University of California, Los Angeles (UCLA) Department of Psychology subject pool and were compensated with course credit.

**Materials and Procedure.** The apparatus was the same as in Experiment 1. The stimuli in Experiment 2 were also the same as previously indicated, except two differences: (1) the motor object was always stationary after impact (i.e., \( v_A = 0 \) cm/sec) and (2) the motor object moved comparatively faster: \( u_A = 6, 11, \) and 15 cm/sec. Instead of using mass ratio, restitution, and each object’s pre-collision velocity to determine their post-collision velocities
in each trial, the ratio of the motor object’s pre-collision velocity to the projectile object’s post-collision velocity (see Figure 3.1C) was directly manipulated across trials: $u_A/v_B = 0.5, 0.7, 1, 1.4, 2$. In addition, a temporal delay (see Figure 3.1B) was placed between the moment of impact and the projectile object’s initial movement: $t = 0, 70, 140, 210, 280$ msec. These manipulations yielded $3$ (motor speed) $\times$ $5$ (velocity ratio) $\times$ $5$ (temporal delay) $\times$ $3$ (background movement) = $225$ collision stimuli presented in a within-subjects design. The trials were presented in a randomized order and no feedback was provided. The experiment lasted approximately $30$ minutes.

Participants began the experiment by viewing a set of instructions informing them that they would be viewing videos of two (equally heavy) discs in motion. Once again, they were informed that there would be vertical grid lines behind the discs that would move leftward/rightward or remain at rest. Following each video, participants were asked, “Did the left object launch the right object?” and responded on a scale from 1 (Definitely No) to 9 (Definitely Yes) with a middle rating of 5 (Unsure).

### 3.4.2 Results

**Human Results.** As in the previous experiment, we averaged individual participants’ ratings across the three trials with different pre-collision motor velocity ($u_A$). Mean ratings in each of the temporal delay, background movement, and velocity ratio conditions are displayed in the top panels of Figure 3.4. A three-way repeated measures ANOVA was conducted on the mean causal ratings with three within-subjects factors. There was a significant two-way interaction between temporal delay and velocity ratio, $F(16, 13) = 2.90, p = .03$, indicating that the impact of velocity ratio on causal ratings depended on the magnitude of temporal delay. The three-way interaction and remaining two-way interactions were not statistically significant.

The impact of velocity ratio on causal ratings was examined in each temporal delay condition. First, we examined the condition without temporal delay ($t = 0$ msec) and found that causal ratings were significantly impacted by velocity ratio, $F(4, 25) = 4.33, p < .01,$
Figure 3.4: Human Data and Model Predictions (Experiment 2) Human (top) and model predicted (bottom) causal ratings averaged across pre-collision motor speed ($u_A$) for each velocity ratio ($u_A/v_B$) condition. Separate plots (left to right) indicate different temporal delay conditions, and separate lines indicate different background (BG) movement conditions. Vertical error bars indicate standard error of the mean.

which replicated Michotte’s original finding that causal perception of the launching effect depends on the ratio between the two objects’ pre- and post-collision speeds. However, when a noticeable temporal delay was introduced, participants rated their causal impression primarily based on the length of the temporal delay—with much less attention given to velocity ratio—as there was no significant simple main effect of velocity ratio in the $t = 70, 140, 210, 280$ msec temporal delay conditions, $F(4, 25) = 2.73, .58, 1.07, 1.36; p = .052, .68, .39, .28$, respectively.

The impact of relative vs. absolute motion on causal ratings of observed launching events was examined in the absence of temporal delay ($t = 0$ msec), because it was in this condition that velocity had an impact on causal ratings. We found that causal ratings, in fact, were
impacted by background movement at a 0 msec delay, $F(2, 27) = 4.02, p = .03$ (see Figure 4, top left panel). Specifically, ratings in the rightward background condition were significantly smaller than ratings in the rest background condition, $F(1, 28) = 7.94, p < .01$, as well as smaller than the leftward background condition, $F(1, 28) = 5.35, p = .03$. These results indicate that when the relative motions of two colliding objects (with respect to a moving background) are slow, people are less likely to report that the motor object launches the projectile object forwards.

**Model Results.** Predictions from the noisy Newton model with absolute velocity inputs are indicated by the black curves in the bottom panels of Figure 3.4. The model predictions were compared with human ratings in the $t = 0$ msec condition because it was here that background movement had a significant impact. The absolute model reveals a fit of $r^2(13) = 0.314$ (95% CI = [0.004, 0.718]). The model’s predictions are not influenced by the background movement condition, so it cannot explain the behavioral result that rightward background movement yielded lower causal ratings than resting and leftward background movement.

However, the causal noisy Newton model with relative motion inputs—as defined in the previous section—can qualitatively explain the impact of background movement on causal perception. Although the model fit with relative velocity inputs is comparable to the absolute velocity model fit, $r^2(73) = 0.376$ (95% CI = [0.013, 0.761]), the predicted causal ratings are systematically influenced by background movement (see separate curves in bottom panels of Figure 3.4). While there was no observable difference between the leftward and rest background predictions, the model shows comparatively lower ratings in the rightward background condition compared with the leftward condition, which is consistent with human results. The model also captures Sanborn et al.’s (2013) finding that a 1/1 velocity ratio corresponds with peak causal ratings. Also note that model ratings achieve floor values at a temporal delay of approximately 210 msec. This occurs because large temporal delays disagree with the prior expectation of a 0 msec delay and thus generate a small temporal likelihood term. Human ratings also appear to reach floor values at around the same tem-
poral delay. We chose the following model parameters in Experiment 2 to account for the human data: $\sigma = 972$, $k_v = .319$, $k_l = .004$, $w_v = .059$, $P(O|NC) = 2 \times 10^{-5}$. Note that a separate set of model parameters were chosen since the range of velocity input values was significantly greater than in the previous experiment.

3.5 General Discussion

The results reported herein demonstrate that (i) the motor object bias in mass ratio judgment is strengthened or attenuated when the background on which colliding objects travel moves either against or with the direction of their motion, respectively; and (ii) impressions of launching are similarly influenced by moving backgrounds when there is no temporal delay between the movements of the two objects. The noisy Newton model (Sanborn et al., 2013) for object collisions was implemented and compared with human data in both a mass ratio judgment and causal rating task. For mass ratio judgment, the model with relative motion inputs accounts for human performance well across a range of experimental conditions. The goodness of the fit suggests that humans use perceived relative motion as the input to high-level cognitive systems when inferring observable physical properties. For causal perception, the model with relative motion inputs explains the finding that background movement with the motion of colliding objects negatively impacts causal ratings.

In summary, our results show that impressions about the attributes of—and relationships between—entities in the world are systematically influenced by low-level spatiotemporal characteristics in observed scenes. However, it remains unclear whether the influence of more abstract contextual properties (e.g., Mayrhofer & Waldmann, 2014) on causal impressions can be explained by their spatiotemporal characteristics alone. Another question is whether human counterfactual reasoning (Gerstenberg et al., 2015, 2017) in object collision tasks are also systematically influenced by background movement. It would be beneficial for future work to explore these possibilities.
 CHAPTER 4

Intuitive Physics in Virtual Reality

4.1 Abstract

This paper examines how humans adapt to novel physical situations with unknown gravitational acceleration in immersive virtual environments. We designed four virtual reality experiments with different tasks for participants to complete: strike a ball to hit a target, trigger a ball to hit a target, predict the landing location of a projectile, and estimate the flight duration of a projectile. The first two experiments compared human behavior in the virtual environment with real-world performance reported in the literature. The last two experiments aimed to test the human ability to adapt to novel gravity fields by measuring their performance in trajectory prediction and time estimation tasks. The experiment results show that: (i) based on brief observation of a projectile’s initial trajectory, humans are accurate at predicting the landing location even under novel gravity fields, and (ii) humans’ time estimation in a familiar earth environment fluctuates around the ground truth flight duration, although the time estimation in unknown gravity fields indicates a bias toward earth’s gravity.

4.2 Introduction

Sending a manned spacecraft to Mars would be a fantastic adventure, yet living on another planet could lead to significant challenges to human perceptual and cognitive systems. The change in gravity itself could alter daily activities (e.g., throwing an object toward a desired location or pouring water into a container) that require adjustment of prediction and action
in light of changing physical properties on the new planet. Imagine that you are living in an environment with a different gravity field than earth. Would you be able to adapt to it quickly? And how accurate would your predictions about the physical world be compared to when you were on earth?

Consumer-level virtual reality (VR) devices, with rapidly increasing popularity, provide a useful means for researchers to conduct experiments that were traditionally too costly or impossible to carry out in the real world. VR allows users to experience an artificial world in a manner similar to how they experience the real world: i.e., head-mounted displays give the impression of three-dimensional observation, and remote controllers afford interactions with the virtual world from an embodied egocentric perspective. In particular, VR technology allows for both the control of many underlying factors of the virtual world—e.g., time (Schatzschneider, Bruder, & Steinicke, 2016) and gravity—and direct measurement of behavioral changes in novel environments.

In this paper, we conducted four experiments to measure human performance in different tasks under novel and familiar gravity fields. In the first two experiments, participants were asked to strike a ball off of a track onto a target location and to trigger a ball to hit a target given a speed rating input. In Experiments 3 and 4, participants were asked to make predictions about the location and flight duration of a projectile given the initial 0.2 seconds of its trajectory. The purpose of the experiments was to examine how humans learn and reason about object motion in novel gravity fields: are humans able to spontaneously habituate to new gravity fields? Do humans implicitly use prior knowledge about earth’s gravity to reason about new environments? Are humans implicitly simulating physical motion or predicting the movements using low-level visual features exclusively?

The first pair of experiments in the present work compare human performance in the VR setting with findings in similar real-world situations (Krist, Fieberg, & Wilkening, 1993). The second pair of experiments compare two types of intuitive physical judgments (location predictions and time estimates) under different gravity fields. In summary, this paper made the following contributions: (i) replicated a previous study on speed production and rating in novel virtual environments to demonstrate that VR is a feasible and reliable tool for
studying human perception and cognition, (ii) carried out a novel experimental design and method which precludes real-world replication, and (iii) measured the effect of gravity field on human behavior in tasks varying in their cognitive demands.

The remainder of this paper is structured as follows: Section 4.3 discusses related work in intuitive physics and virtual reality, Sections 4.5-4.8 describe each experiment’s methods and results, Section 4.9 provides a comparison of results between experiments, and Section 4.10 concludes the study and outlines proposed directions for subsequent work.

4.3 Background and Related Work

4.3.1 Intuitive Physics

The principles of Newtonian physics accurately describe the behavior of objects in our realizable world, yet people’s commonsense beliefs about how objects move are often at odds with ground-truth predictions (Clement, 1982; Kaiser et al., 1986). For example, when reasoning about the trajectories of moving objects on explicit, pencil-and-paper tasks, people commonly predict that an object will follow a curvilinear path upon exiting a C-shaped tube (McCloskey et al., 1980) and that an object dropped from a moving body will follow a linear path downward (McCloskey, 1983; McCloskey et al., 1983) with a velocity proportional to its weight (Shanon, 1976). Although these misconceptions appear consistent with erroneous physical theories (e.g., medieval impetus or Aristotelian principles), people are not internally consistent in their intuitive physical judgments across related tasks at the explicit level, suggesting adherence to domain-specific theories of motion (Kaiser et al., 1986; Yates et al., 1988). More recent work proposes that people are susceptible to erroneous theories of motion in the context of explanation, but adhere to rational inference on the basis of Newtonian physics in the context of prediction: e.g., people are unable to draw the trajectory of an object following release from a pendulum but can successfully place a bucket where they expect the ball to land (Smith et al., 2013). Earlier empirical studies also suggest that humans are less susceptible to erroneous theories of motion when physical situations are pre-
sented in a familiar context (Kaiser et al., 1986) or in an animated format (Kaiser, Proffitt, & Anderson, 1985; Kaiser et al., 1992). In addition, although people are inaccurate when reasoning about certain physical situations (e.g., liquid behavior) explicitly, they respond rationally when repeating similar tasks through simulated action (Schwartz & Black, 1999) or animation-facilitated mental simulation (Kubricht et al., 2016).

The present experiments provide two concrete task domains which rely on action and prediction, rather than explanation (see Figure 4.1). Thus, we hypothesized that participants would adhere to Newtonian (rather than erroneous) theories of motion when reasoning about the physical behavior of moving objects at the implicit level. The tasks, however, differ in their cognitive and sensorimotor demands in addition to the spatial information needed to effectively reason about them. For example, Experiments 1 and 2 (replication of previous work; described in Sections 4.5 and 4.6) provide participants with a perception-for-action and perception-only task, respectively. Although Krist et al.’s original study found no difference between performance in the two tasks (Krist et al., 1993), there is a breadth of evidence which suggests that guided action and perceptual identification rely on two independent neural pathways (Goodale & Milner, 1992). More recent fMRI results, however, indicate that object weight (a non-visual, motor-relevant property) can be represented in regions associated with the (ventral) perceptual identification stream from familiar texture cues (Gallivan et al., 2014). Information about the weight of unfamiliar objects, however, is arguably inferred from haptic feedback, a crucial sensory modality absent in virtual environments that has been pursued rigorously over the past three decades (Burdea, 1999). The effect of the absence of haptic feedback on our perception-for-action task is further discussed in the following sections.

The perception-for-action task in Experiment 1 also differs from Krist et al.’s original task in that it required the use of a tool (i.e., a controller specifying the position of a virtual ball) to propel a second ball off of a platform. Tool use has been shown to directly affect perceived distance, suggesting that people represent the physical world in terms of their ability to interact with it (Witt, Proffitt, & Epstein, 2005). Perception-action recalibration has also been reported in studies on predicted walking distance (real and imagined) following
adaptation (Ziemer et al., 2013), judged hill slant following a loading of weight from a backpack (Proffitt, 2006; but see Durgin, Klein, Spiegel, Strawser, & Williams, 2012), and illusory reversal of temporal order between actions and sensation (Stetson, Cui, Montague, & Eagleman, 2006). Although the potential effect of tool use on Experiment 1’s task is interesting, a more primary aim of the present work is to determine how participants respond and adapt to novel gravitational fields across the four experiments outlined in Sections 4.5-4.8. Are people biased toward believing that the gravitational acceleration of an unknown environment is equal to that on earth? Does this bias manifest itself across tasks differing in their aforementioned cognitive demands? The potential interplay between prior beliefs about gravity and associated human activity is discussed further in Section 4.3.2.

### 4.3.2 Mental Simulation

Previous work has demonstrated that people often employ mental simulation strategies when reasoning about physical situations (Hegarty, 2004; Hegarty & Sims, 1994). For example, expert problem-solvers spontaneously employ mental simulation strategies to anchor assumptions relevant to their explanations when reasoning about a mass-spring system (Clement,
when people are asked to predict the rotation direction of elements in a pulley system, they intuitively simulate motion in an order corresponding to the machine’s causal sequence: i.e., more time is required to reason about motion later in the causal chain (Hegarty, 1992). The time-dependent characteristics of people’s mental simulations suggest spatial (rather than visual) representation, which quantitatively encodes both latent and observable variables relevant to the physical situation (Hegarty, 2004; Schwartz & Black, 1996).

The findings outlined in Section 4.3.1, however, suggest that people’s predictions about one-body motion do not always agree with Newtonian physics. Early work on human judgment in two-body interactions (e.g., the collision of two point masses) reported consistent findings (Runeson, 1983; Todd & Warren Jr., 1982), although judgment biases were subsequently explained by attention to simplified rules or heuristics based on observational cues (Gilden & Proffitt, 1989, 1994; Runeson et al., 2000). Despite the fact that cue-heuristic models explain some human judgment biases qualitatively, several recent models have obtained good quantitative agreement by assuming that people form rational inferences about dynamical systems by combining noisy perceptual inputs with Newtonian physical principles, given prior beliefs about spatially represented variables: i.e., the noisy Newton hypothesis (Sanborn et al., 2013). Following this hypothesis, quantitative judgments about physical systems can be inferred by simulating initial states (sampled from distributions reflecting noisy perception) forward in time using probabilistic Newtonian physics, querying the output states and aggregating judgments across numerous simulations to form predicted response distributions (Battaglia et al., 2013).

Probabilistic simulation models have demonstrated success in predicting human judgments in several domains: e.g., physical scene understanding (Battaglia et al., 2013), object interactions (Liang et al., 2015), liquid dynamics (Bates et al., 2015; Kubricht et al., 2016) and causality in mass-collision displays (Gerstenberg et al., 2015). Their predictions deviate from ground-truth physics in accordance with biases observed in human experiments. Such findings are generally explained by noisy perceptual inference based on prior beliefs about relevant variables in the physical system. For example, the standard belief that an initially
moving object is heavier in a two-body collision is well-explained by a prior belief that objects are more likely to move slow than fast (Sanborn et al., 2013). In the perception-for-action task in Experiment 1, participants must propel a ball to a target location by striking it with their controller. However, they receive no haptic information about the weight of the ball: i.e., the ball feels weightless. Thus, we hypothesized that in the absence of haptic feedback, people will underestimate the force needed to propel a ball to a given target location.

In Experiments 3 and 4 of the present work, participants were asked to reason about the trajectory and flight duration of a projectile moving under different gravity fields (see Sections 4.7 and 4.8). One hypothesis is that people predict future physical states by simulating projectile motion forward in time, holding prior beliefs on the underlying physical variables: e.g., velocity and gravitational acceleration. Given robust experience in earth’s gravity, we predict that participants’ simulations will adhere to a gravitational acceleration biased toward that of earth. Alternatively, people might predict trajectories using more explicit mechanisms based on perceptual identification, such that inferred locations will not be biased toward what would be expected under earth’s gravitational field: i.e., gravitational acceleration—and prior beliefs about its magnitude—will not play into the location prediction process.

4.3.3 Virtual Reality

Virtual reality (VR) technology provides an analog experience in a three-dimensional environment similar to that of the real world. Although the majority of research on VR focuses on the technology itself in order to improve users’ experience of hardware and software—e.g., sensing (Shotton et al., 2011; Weichert, Bachmann, Rudak, & Fisseler, 2013) [39, 43], simulation (NVidia FleX, 2016; Unity 3D Engine, 2016; Unreal engine, 2016) and platform integration (Lin et al., 2016)—recent studies have sought to directly examine human perception and cognition in virtual environments. Previous work has focused on human behavior in situations that are similar between the virtual environment and the real world: e.g., visual perception of egocentric distance in real and virtual environments (Loomis & Knapp, 2003),
and human perception of the three components of locomotion (i.e., distance, speed and time) during immersive walkthroughs (Bruder & Steinicke, 2014). Further work has utilized virtual environments to simulate situations that cannot be emulated in the real world, such as the effect of a naturally or unnaturally moving sun on human time judgments (Schatzschneider et al., 2016), the perception and understanding of the exchange of avatars (Lopez et al., 2014), and the visualization of relativity (Schwartz & Black, 2016). In the present study, two experiments similar to the studies by Krist et al. (1993) were conducted to compare the behavior of humans in a virtual environment to their corresponding behavior in a real world task (see Sections 4.5 and 4.6). Two additional experiments simulated physical situations that preclude real world replication (see Sections 4.7 and 4.8).

4.4 Experiment Overview

**Participants.** A total of 20 participants (8 female and 12 male) participated in the study. Participants were either undergraduate or graduate students at the University of California, Los Angeles. The average age of participants was 22.8 years old with a standard deviation of 2.67. All participants had normal to corrected-to-normal vision.

**Apparatus.** During the experiments, participants wore an HTC Vive head mounted display (HMD) with two 1080×1200 screens (one per eye), a 90 Hz refresh rate, and a 110° field of view. Participants used a native HTC Vive controller to interact with objects and scenes inside the virtual setting. Responses were automatically tracked by the HTC Vive system and recorded by our programs. Two standard HTC Vive base stations (lighthouses) were mounted on the wall to simultaneously track the pose and dynamics (position, velocity, and orientation) of both the HMD and controllers over time. The virtual environment was designed using the Unreal Engine 4 gaming platform, providing state-of-the-art, physics-based simulation in real time.

**Design.** Experiments 1 and 2 were designed to replicate and extend Krist et al.’s original study (1993) in a virtual environment. The design of our experiments and the previous study by Krist and colleagues were identical except that in the present experiment, participants
used an HTC Vive controller to strike a virtual ball (rather than throwing an actual, physical ball) off of a track onto a target location. Participants were free to traverse the virtual environment. Experiments 3 and 4 differed from typical experiment settings in the literature examining human projectile motion predictions (e.g., in Smith et al., 2013). Rather than presenting pre-recorded videos in 2D displays and collecting responses using a keyboard or mouse, we measured participants’ performance in an immersive, 3D environment using a laser beam (measurement tool) in VR. The virtual world provided participants with a vivid and realistic environment that enabled several physical interactions between entities and agents. Furthermore, allowing participants to navigate freely inside the virtual environment provided a means to adjust their individual viewing angle so they could view the entire environment.

**Experiment Order.** Participants were asked to complete 3 blocks of experiments in a within-subjects experimental design. In each block, all four experiments shared the same unique gravity field (i.e., gravity field was manipulated between blocks). The gravity field in each environment was selected from 1.5g, 1.0g and 0.5g. Participants were informed that the gravity field for the first block of experiments would be equal to earth’s gravity (1.0g). In the subsequent two blocks, half of the participants completed the block under half of earth’s gravity field (0.5g) first followed by the block with 1.5 times earth’s gravity field (1.5g); the other half of the participants completed the experiments in the counterbalanced order: i.e., 1.5g first and 0.5g second. Participants were told that they would experience unfamiliar gravity fields in the second and third blocks, but information about the specific gravity field was not provided (i.e., whether the gravity field would be greater than or less than earth’s gravity).

In each block, Experiments 1 and 2 were conducted prior to Experiments 3 and 4 for all participants. To control for order effects, half of the participants completed Experiment 1 prior to Experiment 2, and the other half completed Experiment 2 first. Similarly, half of the participants completed Experiment 3 prior to Experiment 4, and the other half completed Experiment 4 first. The order stayed the same for all three blocks for the same participant.
4.5 Experiment 1: Direct Action

The first experiment asked human participants to propel a ball toward an indicated target location in a virtual environment under different gravity fields. Participants directly interacted with virtual objects in the VR setting. After providing responses using a Vive controller, participants viewed the full trajectories of the propelled objects. We sought to compare human performance in VR with that in the real world (Krist et al., 1993) by examining human performance under a gravity field identical to that on earth (i.e., a familiar gravity field). We further examined how well humans perform under gravity fields different from earth’s gravity (i.e., unfamiliar gravity fields).

4.5.1 Method

**Experiment Setting.** As illustrated in Figure 4.2a, the setting of Experiment 1 consisted of a horizontal track (0.16 meters wide, 0.90 meters long), a red, cubic target (length of each side = 0.12 meters), and a white ball (diameter = 0.08 meters, friction coefficient = 0). In the virtual environment, each controller was represented by a sphere (diameter = 0.10 meters). Participants were instructed to use a Vive controller to hit the white ball on the track onto the red target as indicated in Figure 4.1a. The vertical height of the track and the horizontal distance between the track and the target were chosen from a pre-defined discretized set, identical to Krist et al.’s (1993) previous study. Participants were instructed to hit the ball horizontally so that the ball would exit the track with zero velocity along the vertical axis. They were also informed that only the first collision point of the ball would be counted as a successful hit: i.e., bouncing the ball onto the target or rolling it toward the target would not be counted. The size of the track, the diameter of the ball, and the rendered background environment were held constant across all experiments.

**Training Session.** Participants were first given a demonstration and practice trials to familiarize themselves with the virtual environment. Participants were first shown a 3D visual demonstration through the HTC Vive HMD: a ball leaves the track from a height of 1 meter with an initial horizontal speed of 2 meters per second and lands on the ground.
Figure 4.2: **Experiment Designs.** Notes and measurements overlaid on top of the scenes are for illustration and were not provided to participants. (A) Setting of Experiment 1 in the virtual environment during the testing session: a track at different heights and a red target at different distances were displayed. Participants used a controller to strike the virtual ball on the track onto a red target on the ground. (B) Setting of Experiment 2 during the testing session: similar to Experiment 1, a track and a target were shown to the participants, but no virtual ball was displayed. Participants selected a speed on the slider to indicate at which speed they thought the ball should be projected to hit the target. (C) Setting of Experiments 3 and 4 in the virtual environment: a virtual ball was generated on a platform on the top right side of the display. Once the button on the controller was pressed, the ball was projected horizontally leftward toward the red laser beam. (D) Illustration of Experiments 3 and 4 from the participant’s perspective: participants stand next to the laser beam and perceive a virtual ball flying toward them, which is different from inferring the ball speed from a third-person view.

Next, they were asked to move the controller to hit the ball on the track, which ensured that participants knew how to use the controller prior to the training and testing trials. Finally, participants were instructed to hit the virtual ball with the same initial configuration three times as far as possible and then three times as close as possible.

After demonstration and practice, participants were given training trials: a target appeared on the floor and the height of the track was adjusted. Participants were instructed
to try their best to hit the ball onto the target location. If the ball did not exit the track horizontally in a given training trial, the setting for that trial was presented again at the end of the training session. During the training session, the height of the track was chosen from 0.20, 0.70 and 0.95 meters; the distance between the center of the target and the exit of the track was either 0.30 or 0.90 meters. In total, there were 6 different combinations of distance and height for the training session. During the training session, participants observed the full trajectory of the ball after it left the track (i.e., they received visual feedback in each training trial).

**Testing Session.** After the 6 training trials, participants were presented with 12 testing trials with parameters indicated by the combination of 4 different heights and 3 different distances. Specifically, the height of the track was chosen from 0.20, 0.45, 0.70 and 0.95 meters; the distance between the center of the target and the exit of the track was chosen from 0.30, 0.60, and 0.90 meters. These values were identical to the set of values used in the original study (Krist et al., 1993). Similar to the training session, if participants failed to hit the ball so that it exited the track horizontally in a testing trial, that trial was presented again at the end of the testing session. After hitting the ball, participants were provided with full visual feedback: they viewed the full trajectory of the ball and were informed of whether the ball successfully hit the target.

When participants hit the ball using the controller (represented by a virtual sphere; diameter = 0.10 meters), the speed of the controller was measured by the HTC Vive base station and fed into the Unreal Engine. The Unreal Engine then computed the resulting speed of the ball after the collision using an internal physics engine. The speed of the ball was recorded as the produced speed measurement for each testing trial. To examine whether humans behave rationally, a ground-truth model prediction for each trial was calculated analytically based on the physical parameters (i.e., height, distance, and gravity) in each setting.
4.5.2 Results

ANOVA Results. In order to determine whether the speeds produced by participants depended on the magnitude of gravitational acceleration, the height of the track, and the distance between the track and target, we performed a $3 \times 3 \times 4$ (Gravity $\times$ Height $\times$ Distance) analysis of variance (ANOVA) at the $\alpha = 0.05$ significance level. The three-way interaction term was not significant, $F(12, 684) = 1.34$, $p = 0.189$. There were significant main effects of gravity, $F(2, 684) = 74.05$, $p < .001$, distance, $F(2, 684) = 181.79$, $p < .001$, and height, $F(3, 684) = 40.31$, $p < .001$. There was also a significant two-way interaction between gravity and distance, $F(4, 684) = 4.78$, $p < .001$. However, the interaction between gravity and height was not significant, $F(6, 684) = 1.10$, $p = .360$. Contrary to the ground truth model, the interaction between distance and height was only marginally significant, $F(6, 684) = 1.86$, $p = .086$. This result indicates that the relationship between produced speed and distance did not differ significantly across the implemented track heights. Our results demonstrate that participants’ produced speeds depended on relevant physical factors: i.e., the height of the track, the distance between the track and target, and importantly gravitational acceleration in the virtual world. In addition, the relationship between produced speed and distance also varied according to specific gravity fields, which agrees with the ground-truth relationships. Figure 4.3 depicts the mean speeds produced by participants under three different gravity fields for various height-distance combinations. The three figures qualitatively differ from one another, as evident in the aforementioned Gravity $\times$ Distance interaction.

Regression Analysis. To compare participants’ produced speeds with the ground-truth model predictions, we performed a linear regression analysis as depicted in Figure 4.4. Optimal performance is indicated by a slope of 1.0; a slope less than 1.0 indicates that participants underestimated speed, and a slope greater than 1.0 indicates the opposite. There is a strong linear relationship between speeds produced by human participants and the speeds predicted by the ground-truth model. However, the regression slopes are smaller than 1.0 under each gravity field, indicating that humans move slower than what is optimal given the ground-truth model. Several factors could contribute to the apparent underestimation of speeds in
Figure 4.3: **Human Results (Experiment 1)**. Mean speeds produced by participants in Experiment 1 under three different gravity fields: 1.5g, 1g, and 0.5g. Error bars indicate standard error of the mean.

Figure 4.4: **Regression Results (Experiment 1)**. Mean speeds produced by participants versus ground truth speeds in Experiment 1 under three different gravity fields: 1.5g, 1g, and 0.5g.

Experiment 1’s perception-for-action task. One primary reason could be due to the absence of haptic feedback following interaction with the virtual ball. As the weight of the virtual ball can only be inferred (rather than directly perceived through the sensorimotor system as is typical in real-life situations), participants may have significantly underestimated the weight of the ball. In turn, this may have led participants to underestimate the force needed to propel the virtual ball to hit each target location. Thus, regardless of the gravity field, participants consistently underestimated the speed needed for the task.
Table 4.1: Regression Results (Experiments 1 and 2). Linear relationships between human estimated speeds and speeds predicted by the ground-truth model in Figures 4.3 and 4.5, where $g$ denotes ground truth gravity field ($1.0$ means $1.0g = 9.8$ m/sec$^2$), $h$ (cm) is the height of the track, $s_{gt}$ is the ground truth slope, $s_1$ is the regression coefficient of the data collected in Experiment 1, $\sigma_{s_1}$ is the standard deviation for the produced speeds in Experiment 1, $s_2$ is the regression coefficient for the speed ratings in Experiment 2, and $\sigma_{s_2}$ is the standard deviation of the data collected in Experiment 2.

<table>
<thead>
<tr>
<th>$g$</th>
<th>$h$</th>
<th>$s_{gt}$</th>
<th>$s_1$</th>
<th>$\sigma_{s_1}$</th>
<th>$r_{s_1}^2$</th>
<th>$s_2$</th>
<th>$\sigma_{s_2}$</th>
<th>$r_{s_2}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>20.0</td>
<td>0.0606</td>
<td>0.0251</td>
<td>0.1097</td>
<td>0.9692</td>
<td>0.0537</td>
<td>0.2762</td>
<td>0.9577</td>
</tr>
<tr>
<td>1.5</td>
<td>45.0</td>
<td>0.0404</td>
<td>0.0123</td>
<td>0.0123</td>
<td>0.9522</td>
<td>0.0485</td>
<td>0.1161</td>
<td>0.9905</td>
</tr>
<tr>
<td>1.5</td>
<td>70.0</td>
<td>0.0324</td>
<td>0.0249</td>
<td>0.0249</td>
<td>0.9890</td>
<td>0.0405</td>
<td>0.0916</td>
<td>0.9916</td>
</tr>
<tr>
<td>1.5</td>
<td>95.0</td>
<td>0.0278</td>
<td>0.0211</td>
<td>0.0211</td>
<td>0.9589</td>
<td>0.0344</td>
<td>0.0306</td>
<td>0.9987</td>
</tr>
<tr>
<td>1.0</td>
<td>20.0</td>
<td>0.0495</td>
<td>0.0334</td>
<td>0.2396</td>
<td>0.9212</td>
<td>0.0518</td>
<td>0.2665</td>
<td>0.9577</td>
</tr>
<tr>
<td>1.0</td>
<td>45.0</td>
<td>0.0330</td>
<td>0.0258</td>
<td>0.1919</td>
<td>0.9157</td>
<td>0.0379</td>
<td>0.0243</td>
<td>0.9993</td>
</tr>
<tr>
<td>1.0</td>
<td>70.0</td>
<td>0.0265</td>
<td>0.0170</td>
<td>0.0865</td>
<td>0.9586</td>
<td>0.0357</td>
<td>0.0452</td>
<td>0.9973</td>
</tr>
<tr>
<td>1.0</td>
<td>95.0</td>
<td>0.0227</td>
<td>0.0199</td>
<td>0.2077</td>
<td>0.8468</td>
<td>0.0307</td>
<td>0.0454</td>
<td>0.9964</td>
</tr>
<tr>
<td>0.5</td>
<td>20.0</td>
<td>0.0350</td>
<td>0.0185</td>
<td>0.0836</td>
<td>0.9699</td>
<td>0.0449</td>
<td>0.2081</td>
<td>0.9655</td>
</tr>
<tr>
<td>0.5</td>
<td>45.0</td>
<td>0.0233</td>
<td>0.0159</td>
<td>0.0222</td>
<td>0.9967</td>
<td>0.0401</td>
<td>0.2668</td>
<td>0.9314</td>
</tr>
<tr>
<td>0.5</td>
<td>70.0</td>
<td>0.0187</td>
<td>0.0125</td>
<td>0.0856</td>
<td>0.9280</td>
<td>0.0289</td>
<td>0.0422</td>
<td>0.9965</td>
</tr>
<tr>
<td>0.5</td>
<td>95.0</td>
<td>0.0161</td>
<td>0.0119</td>
<td>0.1157</td>
<td>0.8641</td>
<td>0.0287</td>
<td>0.0936</td>
<td>0.9825</td>
</tr>
</tbody>
</table>

A second linear regression analysis was then performed to examine how produced speed varied as a function of distance for each track height. The regression coefficients, standard errors and the r-squared statistics for Experiment 1 are reported in Table 4.1. The ground-truth slope $s_{gt}$ of the speed-distance relationship was determined by the following expression:

$$s_{gt} = \sqrt{\frac{g'}{2h}},$$  \hspace{1cm} (4.1)

where $g'$ is the gravitational acceleration and $h$ is the track height. For each gravity-height
condition, the human slope was less than the ground-truth slope, providing converging evidence that speed was underestimated regardless of the environment. For Experiment 1, the percent error between the calculated slope $s_i$ and the ground-truth slope $s_{gt}$ for each gravity-height combination was defined by $\frac{|s_i - s_{gt}|}{s_{gt}} \times 100\%$. The mean percent errors (i.e., percent errors averaged across height) for the 1.5g, 1.0g, and 0.5g environments in Experiment 1 were 43.9%, 25.6%, and 34.5%, respectively. The smallest mean percent error was in the 1.0g environment, which agrees with our previous finding where humans performed best in the earth-gravity environment.

**Comparison with Original Study.** The results from Experiment 1 are in general agreement with Krist et al.’s (1993) previous findings. Both studies (ours in VR and Krist’s in the real world), found significant main effects of height and distance, suggesting that humans are sensitive to critical physical variables when interacting with objects, regardless of whether they are virtual or physical. In the virtual environment, however, we found that the interaction effect between height and distance was only marginally significant, whereas the original study (conducted in the real world) revealed a significant interaction between the two factors. The weakened interaction effect was most likely due to participants’ underestimation of produced speed, which may have resulted from the absence of haptic feedback from the VR system. Participants may have developed an implicit bias toward believing the virtual ball was weightless, effectively reducing the variance in their responses and the corresponding power of our statistical analyses.

### 4.6 Experiment 2: Speed Judgment

Although participants were able to move freely in the virtual environment (i.e., all 3D view angles were allowed), they were not able to control or act upon any virtual objects in the present task. The second experiment was designed to examine the human ability to estimate the initial speed of a ball required to hit a target location under different gravity fields. In this task, participants were not asked to perform an action: i.e., striking a ball. Instead, they were asked to give a speed rating.
4.6.1 Method

Experiment Setting. Experiment 2 employed the same track, ball, and target as in Experiment 1. Objects in Experiment 2, however, were stationary: i.e., instead of allowing participants to directly interact with the virtual objects, a slider was introduced to gauge human participants’ speed ratings (see Figure 4.2b). For each stimulus, participants estimated the initial speed of the ball required to hit the target on the ground. Participants were asked to report their estimated speed using a slider. The leftmost side of the slider represented the slowest speed (0.1 meters per second), and the rightmost side of the slider represented the fastest speed (5.5 meters per second). In the experiment, participants used a Vive controller to move the slider to indicate their estimated speed. The reading of the slider was converted into speed by the following expression:

\[ s = p \times (s_{\text{max}} - s_{\text{min}}) + s_{\text{min}}, \]

(4.2)

where \( s \) is the horizontal speed of the ball, \( p \) is the reading of the slider (between 0 and 1), and \( s_{\text{max}} \) and \( s_{\text{min}} \) are the maximum and minimum speeds, respectively.

Training Session. At the beginning of each experiment, three demonstration trials were provided to participants. In each demonstration, a ball traveled along the track at three different speeds: maximum speed (5.5 meters per second), minimum speed (0.1 meters per second), and medium speed (2.8 meters per second). The speed was indicated on the slider at each corresponding position and was visible to the participants.

After observing the three demonstrations, identical settings with the same three predefined speeds were shown again but in a random order. This time, the speed of the ball was not explicitly provided to the participants. Instead, the participants were asked to move the slider using their controller to indicate the corresponding speed. If participants did not answer correctly, they were asked to repeat the trial. If participants failed the second time, they did not proceed to the next part of the experiment. In the real experiment, everyone answered this part correctly after the first or second time. Therefore, every participant pro-
ceeded to the next part of the experiment. Similar to Experiment 1, participants completed 6 different training trials, one for each distance-height combination. The task, however, differed in that participants triggered the ball’s movement by indicating their speed rating on a virtual slider rather than striking the virtual ball with their controller. In each of the training trials, the trajectory of the ball was displayed after leaving the horizontal track.

**Testing Session.** In the testing session, a track and a target were shown at 12 different height-distance combinations (same combinations as Experiment 1). However, in Experiment 2, after participants selected a speed on the slider, the trajectory of the ball was not displayed after leaving the horizontal track. Hence, no visual feedback was provided to the user by the VR system in Experiment 2. This was done to minimize rapid learning based on previous testing trial outcomes.

### 4.6.2 Results

**ANOVA Results.** We carried out the same analysis in Experiment 2 as in Experiment 1. We first performed a $3 \times 3 \times 4$ (Gravity $\times$ Height $\times$ Distance) ANOVA at the $\alpha = 0.05$ significance level to determine whether participants’ speed ratings depended on the magnitude of gravitational acceleration, the height of the track, and the distance between the track and target. The three-way interaction term was not significant, $F(12,684) = 0.53$, $p = 0.896$. Results from the ANOVA indicate significant main effects of gravity, $F(2,684) = 69.35$, $p < .001$, height, $F(2,684) = 787.81$, $p < .001$, and distance, $F(3,684) = 105.08$, $p < .001$, which agrees with Experiment 1’s results. We found significant two-way interactions between gravity and distance, $F(4,684) = 4.61$, $p = .001$, but the effect of gravity did not interact with height, $F(6,684) = 0.60$, $p = .733$. Interestingly, we found a significant Height $\times$ Distance interaction in Experiment 2, $F(6,684) = 14.06$, $p < .001$, which agrees with the ground-truth model and previous findings in the real-world environment (Krist et al., 1993). Figure 4.5 depicts the mean estimated speed for different distance-height combinations under three different gravity fields. The influence of speed and distance on human speed estimation appears to qualitatively vary across the four track heights, as evident in the significant
Figure 4.5: **Human Results (Experiment 2)**. Mean estimated speed as a function of distance between the track and the target in Experiment 2 under three different gravity fields: 1.5g, 1g, and 0.5g. Error bars indicate standard error of the mean.

Distance × Height interaction. Furthermore, the relationship also appears to vary across the three gravity fields, as evident in the significant Gravity × Height interaction.

Running a random effects ANOVA with two-way interactions, we found there was not a significant effect of participant on speed rating, $F(19, 544) = 1.59, p = .083$. There were significant interaction effects between gravity and participant, $F(38, 544) = 5.57, p < .001$, distance and participant, $F(38, 544) = 4.34, p < .001$, and height and participant, $F(57, 544) = 1.91, p < .001$. These findings suggest larger individual differences in perceived gravity, distance, and height in the speed judgment task in Experiment 2 compared with the direct action task in Experiment 1.

**Regression Analysis.** We examined the linear relationship between participants’ speed ratings and the ground-truth speed as depicted in Figure 4.6. The regression analysis again shows a strong linear relationship between the two speeds under each gravity field. The slope under earth’s gravity is 1.0, indicating that participants were highly accurate when triggering a ball to move toward a target location in a familiar environment. The slope in the 1.5g environment is less than 1.0, and the slope in the 0.5g environment is greater than 1.0, indicating that participants underestimated and overestimated speed in each respective environment when action was not involved. These findings suggest that participants’ beliefs
about gravitational acceleration in the 1.5g and 0.5g environments were biased toward earth’s gravity field when they were asked to mentally estimate the speed rather than physically performing the action. Note that humans underestimated speed under all gravity fields in Experiment 1. However, in Experiment 2, humans’ rated speeds under unfamiliar gravity fields showed a strong bias toward earth’s gravity and even showed a slope of 1.0 under earth’s gravity. The discrepancy between the two experiments is likely due to more implicit reasoning involved in Experiment 1’s direct action task and more explicit reasoning based on low-level physical knowledge in Experiment 2’s speed judgment task.

Next, we performed a linear regression analysis to quantify how speed rating varies as a function of distance under each gravity field. Calculated regression coefficients and their corresponding standard errors for Experiment 2 are reported in Table 4.1. The mean percent error between the ground-truth slope and the human slope from the regression analysis was 20.0%, 22.4%, and 58.3% for the 1.5g, 1.0g, and 0.5g environments, respectively. The speed-distance slopes for all track heights in the 0.5g environment were greater than the corresponding ground-truth slope, suggesting a bias toward earth’s gravity. In the 1.5g environment, however, the speed-distance slope exceeded the ground-truth slope for three of the four track heights. This appears to indicate a bias away from (rather than toward) earth’s gravity field, which disagrees with results from the regression analysis comparing
human speed ratings to ground-truth predictions in Experiment 2. However, this discrepancy needs to be interpreted cautiously since humans showed much larger variability in the 1.5g environment. Specifically, participants were increasingly inconsistent when the track was rendered near to the ground in the 1.5g environment, as evident in the large standard errors on the corresponding regression coefficients.

**Comparison with Original Study.** Results from Experiment 2 are in agreement with Krist et al.’s (1993) previous findings. The ANOVA results for participants’ speed ratings showed the expected results, including significant main effects of height and distance and a significant interaction between the two variables. However, in the explicit reasoning task, we also noticed a strong bias toward earth’s gravity field, which suggests the use of low-level, common-sense physical knowledge that over-generalized to novel situations.

### 4.7 Experiment 3: Contact Location Prediction

Our third experiment was designed to examine the human ability to predict the contact location of a projectile’s trajectory under familiar and unfamiliar gravity fields. The ball’s trajectory was briefly displayed and then occluded prior to measuring participants’ predictions. Thus, participants were required to extrapolate projectile motion according to limited visual input. The aim of the present experiment was to determine the reasoning strategies people employ when predicting future projectile locations: do people propagate spatially represented objects forward in time using a mental simulation engine, or do they rely on more explicit reasoning strategies?

#### 4.7.1 Method

**Experiment Setting.** As illustrated in Figure 4.2c, the virtual environment in Experiments 3 and 4 consisted of a tilted laser beam, a launching platform suspended in the air (height = 3 meters), and a white ball (diameter = 0.08 meters, friction coefficient = 0) resting on top of the platform. The angle between the laser beam and the ground was 45°, and the
horizontal distance between the bottom of the laser beam and the platform was 3 meters. In the experiment, the white ball moved horizontally with a random initial velocity, and the ball disappeared 0.2 seconds after leaving the platform. Participants were asked to predict the location on the laser beam where they believed the ball would make contact. The trajectory of the ball always intersected with the laser beam. The reason for choosing a specified orientation for the laser beam was to ensure that participants accounted for gravity when estimating the flight duration of the ball in Experiment 4 (see Section 4.8).

**Training Session.** At the beginning of each experiment, participants were shown one full trajectory of a ball moving from the launching platform to the contact location on the laser beam. A second ball was then presented with the same initial speed but disappeared 0.2 seconds after leaving the platform. Participants were then asked to use their controller to indicate where the ball would make contact with the laser beam. The training session was designed to familiarize participants with the controller and the task procedure. Participants were not provided with any feedback on the true contact location nor the accuracy of their decisions.

**Testing Session.** In the testing session, participants were presented with 10 testing trials in a randomized order. In each trial, the ball moved with a different initial speed and disappeared 0.2 seconds after leaving the platform. Participants were then asked to predict the contact location on the laser beam. The location indicated in the virtual environment served as the location prediction measurement for each participant. No feedback was given following each response.

The experiment was conducted under three different gravity fields (1.5g, 1.0g, and 0.5g). The initial speed of the ball, $s$, for each trial was calculated using the following expression:

$$s = \frac{\tan(\pi/2) \times h}{\sqrt{2h/g'}},$$  \hfill (4.3)

where $g'$ is the gravitational acceleration and $h$ is the height of the contact location on the laser beam. Height was selected from 10 different values: 1.07, 1.17, 1.25, 1.31, 1.37, 1.42, 1.48, 1.54, 1.62, and 1.72 meters. These values were chosen from a Gaussian distribution.
such that the true contact points were denser in the middle and sparser on both ends of the laser beam. Experiments under different gravity fields shared the same set of heights but in different (randomized) orders.

4.7.2 Results

ANOVA Results. We conducted an ANOVA on the percent error \( \left( \frac{|H_h - H_{gt}|}{H_{gt}} \times 100\%\right) \) between participants’ predicted contact locations \( H_h \) and the corresponding ground-truth value \( H_{gt} \) for each height condition. Results from the ANOVA indicate that the error was not significantly influenced by different gravity fields, \( F(2,597) = 0.33, p = 0.717 \). There was a significant influence of height on the percent error, \( F(9,597) = 4.34, p < .001 \).

Trajectory Models. To determine how participants predicted the end point of the trajectory, we compared human performance to four different geometric models. Each model served as a separate hypothesis for predicting the trajectory contact location. Human predictions were compared to each candidate hypothesis:

- **Linear.** The linear model served as a baseline model with the simplest form of contact location prediction.

- **Parabola under current gravity.** The parabola under current gravity model provided the ground-truth contact location for the trajectory in the current environment: 1.5g, 1.0g, or 0.5g.

- **Parabola under Earth gravity.** Considering that participants might have had a bias toward earth’s gravitational acceleration, we compared each prediction to the contact location for the parabolic trajectory in the earth environment: 1.0g.

- **Circle.** Considering that people have rich experiences with circular motion in daily life, one possible hypothesis is to interpret the observed trajectory components as part of circular object movement. We compared human predictions to the contact location for a circular trajectory.
**Trajectory Model Results.** To test the candidate models, we fit each trajectory to sampled points from the initial 0.2 seconds of the trajectory and computed the mean squared error (MSE) between each model’s predicted contact location and human responses. To fit each model, we sampled 10 equally spaced points from the first 0.2 seconds of the observed trajectory. Model parameters were then computed to fit the 10 sampled points such that the MSE was minimized. For the circle model, the least squares method determined a local MSE minimum for the circle’s center and radius. The initialization of the parameters influenced the estimated results, so we swept through 20 different circle centers and 30 different radii (600 parameter combinations) for the initialization and picked the parameters that corresponded with minimum MSE.

Figure 4.7 depicts the contact locations predicted by each candidate trajectory model and human contact location predictions in each environment. Results indicate that humans are remarkably accurate at predicting contact locations given the initial 0.2 seconds of a projectile’s trajectory. Comparing predictions from the four candidate models, the *parabola under current gravity* (ground-truth) model provided the best quantitative fit to human contact location predictions under each gravity field (see Table 4.2). The average MSE across environments was approximately 10 centimeters, which is fairly accurate given that the cross-section of the Vive controller—which participants used to indicate their contact location predictions—was 11.7 centimeters × 8.3 centimeters. There was no observed bias toward the *parabola under earth gravity* model in either of the unfamiliar environments (i.e., 1.5g and 0.5g). The present analysis shows that humans can successfully predict future trajectory locations based on limited visual input, and this ability is not hindered in novel physical environments with non-earth gravity fields.

Note that Experiment 3 was different from Experiments 1 and 2 in both the visual inputs provided to participants (e.g., the laser beam, platform, etc.) and its corresponding task demands. Recall that the first two experiment settings were always presented to participants prior to Experiments 3 and 4. Thus, the experimental design made it possible for participants to generalize knowledge about gravity from the first two experimental settings to later experiments since they were informed that all four experiments in each block shared
Figure 4.7: Human and Model Results (Experiment 3). Contact locations predicted by the four candidate trajectory models in Experiment 3. The black dots on the top right corner indicate the 10 sampled points from the initial 0.2 seconds of each trajectory. The oblique line represents the laser beam, and the horizontal line represents the ground. The red dot on the laser beam indicates participants’ mean contact location predictions. Each row depicts results for a different environment: 1.5g, 1.0g, and 0.5g (top to bottom). Each column indicates a different platform height: 1.07, 1.31, 1.48, and 1.72 meters (left to right).

the same gravity field. However, we found that participants showed a global bias toward contact locations predicted by the linear trajectory model. If participants employed a prior belief that the gravity field in an environment should correspond with that on earth, one would expect a bias toward the contact locations predicted by the parabola under earth gravity model in each of the unfamiliar environments. This result suggests that participants may have employed explicit, perceptual knowledge (e.g., spatial location and path geometry) when making their contact location predictions (i.e., prior beliefs about gravitational acceleration did not appear to weigh into participants’ contact location predictions).
Table 4.2: **Model Results (Experiment 3)**. Mean squared error (in meters) of each candidate trajectory model in Experiment 3.

<table>
<thead>
<tr>
<th>Gravity \ Model</th>
<th>Circle</th>
<th>Line</th>
<th>Parabola</th>
<th>Earth Parabola</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5g</td>
<td>0.88</td>
<td>1.14</td>
<td>0.07</td>
<td>0.46</td>
</tr>
<tr>
<td>1.0g</td>
<td>0.90</td>
<td>1.26</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>0.5g</td>
<td>0.98</td>
<td>1.51</td>
<td>0.10</td>
<td>1.03</td>
</tr>
</tbody>
</table>

### 4.8 Experiment 4: Flight Duration Estimation

The fourth experiment was designed to study the human ability to estimate the flight duration of a projectile under familiar and unfamiliar gravity fields. Unlike in Experiment 3, the purpose here was to analyze human time estimation (rather than spatial location prediction) given occluded projectile motion. The aim of the present task was to determine the reasoning strategies people employ when estimating the duration of physical events: do people estimate flight duration using explicit reasoning strategies, as suggested in Experiment 3? Alternatively, do people rely on implicit reasoning strategies (e.g., mental simulation) when performing the temporal task?

#### 4.8.1 Method

**Experiment Setting.** The design of Experiment 4 was similar to that of Experiment 3. However, instead of predicting the contact location of the projectile, participants were instructed to click the trigger on their controller when they believed the occluded ball made contact with the laser beam. The laser beam in Experiment 4 was tilted at a 45° angle, identical to the previous experiment. The laser beam was tilted to ensure participants accounted for gravity when estimating flight duration: i.e., if the laser beam was horizontal, the flight duration would remain constant across trials in a given environment. Moreover, if the laser beam was vertical, flight duration would only depend on the horizontal velocity of the projectile.
Training Session. At the beginning of each experiment, participants were presented with one full trajectory of the ball with an unknown initial speed. Participants were instructed to click the controller once the ball made contact with the laser beam. If participants clicked more than 0.05 seconds earlier or later than the true contact time, they were asked to repeat the trial with the same initial speed until they responded within the 0.05 second window. In the second practice trial, participants were presented with another ball with the same initial velocity and were asked to click the trigger when they thought the ball contacted the laser beam. This time, participants were not given feedback, and the ball disappeared 0.2 seconds after leaving the platform. Similar to Experiment 3, the training session was designed to familiarize participants with the controller and flight duration estimation task. The initial speed used in the training session was not observed in the testing session.

Testing Session. In the testing session, participants were presented with 10 testing trials with different initial speeds and were asked to click the button on the controller once the ball made contact with the laser beam. The ball disappeared 0.2 seconds after leaving the platform in each testing trial. The heights of the contact locations were chosen from the same set of heights as in Experiment 3. After each prediction, participants were not provided with feedback regarding the ground-truth flight duration nor their accuracy. Flight duration measurements for each participant were determined by subtracting the time the ball left the platform from the response time indicated by participants on their controllers. The experiment was conducted under the three different gravity fields employed in the previous experiments: 1.5g, 1.0g, and 0.5g.

4.8.2 Results

ANOVA Results. In order to examine the effect of gravity on participants’ flight duration estimates, we conducted an ANOVA on participants’ flight duration errors (i.e., the difference between participants’ estimated flight durations and the ground-truth flight durations) in each trial. Results from the analysis revealed a significant main effect of gravity, $F(2,597) = 6.99, p = 0.001$. The present results indicate that participants accounted for gravitational
Table 4.3: Human Results (Experiment 4). Difference between mean flight duration estimates and the ground-truth (earth-gravity) flight duration (in ms) in Experiment 4. For each environment (1.5g, 1.0g, and 0.5g), the first row indicates the difference between participants’ mean flight duration estimates and the ground-truth flight duration. The second row indicates the difference between participants’ mean flight duration estimates and the flight duration under earth’s gravity. Positive and negative values indicate overestimation and underestimation, respectively. Height is in meters.

<table>
<thead>
<tr>
<th>Model</th>
<th>1.72</th>
<th>1.62</th>
<th>1.54</th>
<th>1.48</th>
<th>1.42</th>
<th>1.37</th>
<th>1.31</th>
<th>1.25</th>
<th>1.17</th>
<th>1.07</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5g</td>
<td>42.9</td>
<td>52.2</td>
<td>22.6</td>
<td>26.0</td>
<td>38.6</td>
<td>-7.6</td>
<td>-25.0</td>
<td>15.0</td>
<td>17.1</td>
<td>32.9</td>
<td>21.5</td>
</tr>
<tr>
<td>—</td>
<td>15.4</td>
<td>21.7</td>
<td>-10.4</td>
<td>-9.0</td>
<td>1.7</td>
<td>-46.2</td>
<td>-65.6</td>
<td>-27.6</td>
<td>-28.2</td>
<td>-15.8</td>
<td>-16.4</td>
</tr>
<tr>
<td>1.0g</td>
<td>38.7</td>
<td>43.5</td>
<td>10.9</td>
<td>14.7</td>
<td>55.3</td>
<td>-38.1</td>
<td>-16.3</td>
<td>-3.1</td>
<td>-41.4</td>
<td>-33.5</td>
<td>3.1</td>
</tr>
<tr>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>0.5g</td>
<td>25.8</td>
<td>-10.7</td>
<td>-47.4</td>
<td>-36.4</td>
<td>-12.2</td>
<td>-39.9</td>
<td>-90.6</td>
<td>-99.0</td>
<td>-31.2</td>
<td>-46.7</td>
<td>-38.8</td>
</tr>
<tr>
<td>—</td>
<td>123.3</td>
<td>95.6</td>
<td>65.8</td>
<td>82.1</td>
<td>111.6</td>
<td>88.2</td>
<td>42.7</td>
<td>39.5</td>
<td>114.2</td>
<td>107.4</td>
<td>87.0</td>
</tr>
</tbody>
</table>

acceleration when reasoning about the flight duration of an occluded projectile. Table 4.3 provides participants’ mean flight duration errors across height and gravity conditions. As indicated in the rightmost column of the table, participants’ flight duration estimates were biased toward flight durations under earth’s gravity: i.e., flight duration estimates were over- and underestimated in the 1.5g and 0.5g environments, respectively. Thus, participants’ flight duration estimates were biased toward earth-gravity flight durations in the unfamiliar (1.5g and 0.5g) environments.

4.9 Comparison Between Experiments

Comparison of Experiments 1 and 2. In both Experiments 1 and 2, participants took into consideration all three experimental parameters (i.e., gravity, height, and distance) when propelling a ball off of a track onto a target location. The ANOVA in both experi-
ments showed interaction effects between gravity and distance but did not show a significant interaction between gravity and height. One interesting difference between the results of Experiments 1 and 2 is that the ANOVA on produced speed in Experiment 1 showed a marginally significant interaction between distance and height, while the ANOVA for Experiment 2 reported a significant result in agreement with previous findings (Krist et al., 1993). As mentioned previously, this may have been due to participants significantly underestimating the weight of the projectile ball due to the absence of haptic feedback in Experiment 1. Alternatively, the distance in Experiment 1 may have been underestimated in each environment due to the use of a tool (i.e., a Vive controller). This would be consistent with previous reports that tool use can reduce perceived distance (Witt et al., 2005).

Comparing the present results with those from the adult group in Krist et al.’s study, we found that the distance versus speed rating relationship qualitatively agrees between the VR experiment and previous work in the real-world situation: both experiments revealed a strong linear relation between speed estimates and distance, and people’s speed ratings varied across different levels of height. However, we found that participants’ produced speeds were slower than they should have been according to the ground-truth physical model in each environment, leading to a somewhat nonlinear trend between produced speed and distance. In Krist et al.’s original study, participants physically pushed a ball along a track to propel it toward indicated target locations (Krist et al., 1993). Thus, participants could adjust their force input—and associated produced speeds—during the testing phase to match their ideal target speed. In the present experiment, participants hit a virtual ball with a second ball (corresponding to the Vive controller’s location) in an instant and received no haptic feedback: i.e., they were missing an informative variable in their perceptual-motor representation. This lack of sensorimotor feedback might have made it harder for participants to monitor and adjust the magnitude of their input in real time and perhaps caused them to produce speeds that were biased toward a “moderate” magnitude. This uncertainty in sensory input might have given rise to the nonlinear relationship between produced speed and distance observed in the virtual environment in Experiment 1.
Comparison of Experiments 3 and 4. To compare participants’ contact location predictions and flight duration estimates in the last two studies, we further inferred the height of the contact location of the ball according to the flight duration estimates in Experiment 4 using the following expression:

\[
H_{\text{infer}} = 3 - 0.5 \times g' \times t^2, \tag{4.4}
\]

where \(H_{\text{infer}}\) is the inferred height of the contact location, 3 is the initial height (in meters) of the ball, \(g'\) is the gravitational acceleration in the environment, and \(t\) is the flight duration estimate for each participant. The mean difference, \(\delta_{H_{\text{infer}}}\), between the inferred height and the ground-truth contact location across trials is presented in Table 4.4. In each environment, the inferred contact location error from Experiment 4 was at least double the contact location error measured in Experiment 3. If people were using the same reasoning strategy in both tasks, one would expect equivalent magnitudes of error. This suggests that participants did not employ the same reasoning strategy in the contact location prediction (spatial processing) and flight duration estimation (temporal processing) tasks. Hence, when examining and comparing human behavior across various intuitive physical tasks, it is important to carefully examine the specific type of processing involved.

Table 4.4: Human Results Comparison (Experiments 3 and 4). Mean error of inferred contact location (converted from flight duration estimates in Experiment 4) and predicted contact location (from the results of Experiment 3) in each environment.

<table>
<thead>
<tr>
<th>Gravity</th>
<th>(\delta_{H_{\text{infer}}} ) (m)</th>
<th>(\delta_{H_{\text{predict}}} ) (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5g</td>
<td>0.26</td>
<td>0.07</td>
</tr>
<tr>
<td>1.0g</td>
<td>0.24</td>
<td>0.12</td>
</tr>
<tr>
<td>0.5g</td>
<td>0.31</td>
<td>0.10</td>
</tr>
</tbody>
</table>
4.10 Conclusion and Future Work

The present work examined human performance in two projectile motion paradigms using state-of-the-art VR technology to produce artificial physical scenarios. We assessed human performance under natural and unnatural gravity fields in replicated (Krist et al., 1993; Experiments 1 and 2) and novel (Experiments 3 and 4) settings and systematically examined gravitational effects on human performance. Results in the virtual environment from Experiments 1 and 2 were qualitatively in agreement with previous findings in the real world, although the linear relationship between speed and distance was more pronounced in the speed rating (perception-only) task compared to the direct action (perception-for-action) task. Results demonstrate a strong relationship between produced/rated speeds and target distance, and this linear relationship varied across gravity fields (in accordance with the ground-truth relationship). This indicates that participants consistently attended to gravitational acceleration when producing and rating speeds. Mean speed rating errors were negative in the 1.5g setting and positive in the 0.5g setting, indicating that participants’ representations of gravitational acceleration were biased toward earth’s gravity in our perception-only task. This finding reinforces the hypothesis that people infer the physical behavior of their environment by combining noisy perceptual inputs with Newtonian principles given prior beliefs about represented variables (e.g., gravity; Smith et al., 2013). The present results provide evidence that humans hold a strong prior belief about gravitational acceleration which appears distributed around earth’s gravitational acceleration ($\approx 9.8m/s^2$) and leads to apparent biases in their physical intuitions.

The tasks in Experiments 1 and 2 appear drastically different in their cognitive demands, as evident in the qualitative differences between their speed-distance relationships: i.e., the nonlinear trend in Experiment 1 deviates from the ground-truth model and previously reported findings (Krist et al., 1993). Furthermore, research in intuitive physics would predict superior performance in the speed production task due to its concrete task domain (Kaiser et al., 1986; Kaiser, Proffitt, & Anderson, 1985; Kaiser et al., 1992; Smith et al., 2013). One key difference in the speed production task, however, is the role of motor input and
haptic feedback in producing desired projectile speeds for each experimental trial. In the VR environment, no haptic feedback was provided following propulsion of the ball. Thus, the strength of each hit (manifested in participants’ perceptual-motor representations) was inferred rather than directly perceived as is generally the case in daily life. This discrepancy introduces additional uncertainty into the physical environment, which in turn biases the inferred forces toward some prior belief or expectation. Based on these findings, future work in VR environments should work toward providing haptic feedback about experienced forces (perhaps by administering vibrations of variable intensity or other low-level sensory cues; see Burdea, 1999) in order to create an environment with multisensory input that is more consistent with the real world.

Participants were reasonably accurate, however, when reasoning about future trajectory locations in the subsequent experiments, where they were no longer required to provide motor input to the virtual environment. Furthermore, we found that participants’ time judgments were biased toward those that would be expected under earth’s gravity, although their location predictions were remarkably accurate. Unlike their flight duration estimates, participants’ location predictions did not vary significantly across gravitational fields. Our results imply that participants may infer flight durations by reasoning about each trajectory outside of perception. This strategy was mentioned in subjective reports from some participants that they formed their time estimates by adjusting their gaze according to where they thought the ball should be) following occlusion. This strategy agrees with results from physical simulation models, where future physical states are propagated forward in time, given prior beliefs about observable and hidden variables (Bates et al., 2015; Battaglia et al., 2013; Gerstenberg et al., 2015), specifically gravity in the present case. Similar to findings from Experiments 1 and 2, responses were once again biased toward what would be expected under earth’s gravitational field.

However, participants did not appear to adopt a simulation approach to reason about the locations at which the balls contacted the laser beam. Since simulations occur in real time, using a simulation heuristic to predict a future location would be pointless: i.e., future physical states would not be determined until they actually happened. Naturally, a baseball
player catching a fly-ball does not need much time to move to where he predicts the ball will land. Instead, the location prediction task appears explicit, relying on perceptual cues and prior knowledge that objects tend to move along parabolic trajectories under gravity. Participants in the present study upheld this belief, although their responses were consistently biased (but only slightly) toward the linear trajectory. This agrees with previous findings in the intuitive physics literature, where people can accurately predict the end location of a trajectory, although they are less accurate when explaining the trajectory on pencil-and-paper tasks (Smith et al., 2013).

Taken together, the results in the present work demonstrate that humans maintain an impressive ability to habituate to novel physical environments and appear unhindered in predicting future locations of observed trajectories across varying gravitational fields. Our results suggest that humans on a mission to Mars would have minimal difficulty adapting to new gravities, and we conjecture that people’s biases toward Earth’s gravitational field would diminish over time through learning. We also suspect that this bias would be lessened in the real world since people would feel the weight of their bodies change across environments. Future work should aim to further explore this bias, perhaps by weighing down participants’ bodies in gravity fields greater than earth’s.

Our experiments provide a set of physical reasoning problems that lend themselves to varying degrees of spatial representation: i.e., participants appeared to represent latent gravitational acceleration in Experiments 1, 2, and 4 but relied on observable position and velocity in Experiment 3. Such representations can be modeled according to noisy human perception, and future work should aim to determine whether associated probabilistic simulation approaches match well to behavioral measurements. In addition, the present work—unlike Krist et al.’s (Krist et al., 1993) previous study—did not explore performance across age groups. It would be interesting for future work to determine how different age groups (e.g., children and the elderly) habituate to novel virtual gravity fields.
CHAPTER 5

Reasoning Through Mental Simulation

5.1 Abstract

The physical behavior of moving substances is highly complex, yet people can interact with them in their everyday lives with ease and proficiency. To investigate how humans achieve this remarkable ability, the present study examined human performance on an extension of the classical water-pouring problem (Schwartz & Black, 1999) and a substance dynamics prediction task adapted from previous work (Bates et al., 2015). Across three experiments, participants were tasked with (i) judging the relative pouring angle of two substance-filled containers which varied in the volume and viscosity of their contents, (ii) predicting the resting geometry of sand pouring from a funnel onto a surface, and (iii) predicting the dynamics of three substances—liquid, sand, and sets of rigid balls—flowing past obstacles into two basins. Our findings indicate that people do not rely on simple qualitative heuristics based on substance properties (i.e., viscosity, friction, and ball restitution) or perceptual variables (i.e., substance volume and position) when forming judgments and predictions. Instead, computational results from an intuitive substance engine (ISE) model employing probabilistic simulation support the hypothesis that humans infer future states of perceived physical situations by propagating noisy representations forward in time using approximated rational physics. The ISE model outperforms ground-truth physical models in each experiment, as well as competing data-driven learning approaches. Our results expand on previous work proposing usage of mental simulation in physical reasoning and demonstrate human proficiency in predicting the dynamics of sand, a substance that is less common in daily life than liquid or rigid objects.
5.2 Introduction

Imagine that you are a server at a restaurant carrying dishes and beverages from a kitchen
to customers tables. To do this, you must decide where to hold the dishes and at what
orientation to prevent their contents from spilling. More impressively, you must achieve
this while navigating through tables, chairs, and customers in the environment. In the
unfortunate case that a substance-filled container (e.g., a glass of water or a bowl of soup)
topples over and spills onto an occupied table, you must also decide where the substance will
travel and how it will interact with obstacles resting on the surface, hoping to intercept it
before pouring onto an unfortunate customer. We encounter similar situations frequently in
our daily lives while interacting with non-solid substances ranging from granular materials
(e.g., sugar, salt, or sand) to viscous liquids (e.g., syrup or honey), contained in receptacles of
various shapes and sizes. How is the human cognitive system able to rapidly form predictions
and judgments about the dynamics (i.e., physical motion) of substances to allow for such
interactions?

Over the past several decades, the field of intuitive physics has examined the human ca-
pacity for perceiving and reasoning about situations in the physical world. These studies have
primarily explored predictions and judgments that people make about rigid bodies—e.g., the
path of a moving projectile or the relative weight of colliding objects (see Kubricht et al.,
2017 for a review)—rather than non-solid substances. Although comparatively fewer studies
have explored human reasoning about the dynamics of substances, their results have given
rise to quite perplexing results. For example, the Piagetian water-level-task (WLT; Piaget
& Inhelder, 1956; Rebelsky, 1964) was originally designed to determine at what develop-
mental period children begin to represent and attend to horizontal referents in Euclidean
space. In the task, an empty 2D container is rotated and displayed on a piece of paper,
and participants are instructed to draw the surface of the contained liquid, given that it
intersects a specified location on the containers inside surface (see Figure 5.1A). The correct
response is to draw a horizontal line which lies parallel to the bottom edge of the paper,
but approximately 40% of adults draw water lines that deviate from the horizontal by 5
Figure 5.1: **Illustration of Water Tasks.** Stimulus images from (A) the water-level task (WLT) and (B) the water-pouring problem (WPP) are shown. (A) The solid line indicates the correct response where the surface of the contained water is horizontal; the dashed line indicates an incorrect response where the liquid’s surface is rotated 5° from the correct position, in the direction of container rotation. (B) In the WPP, the correct response is that the thinner container should spill last if rotated to the same degree as the wider container.

degrees or more (McAfee & Proffitt, 1991). This appears to suggest that humans do not understand that liquid surfaces should remain horizontal regardless of the orientation of their containers, although this finding has generally been attributed to mental representation of substance position relative to a rotated frame of reference fixed to the container (see McAfee & Proffitt, 1991). In other words, since there are no spatial referents near the container images to gauge rotation (e.g., a line indicating the ground or a vertical wall), the container edges themselves must be used as the “coordinate axes” for the represented situation. This perspective is reinforced by research showing that employees in professions where containers are interacted with regularly (e.g., waitresses and bartenders) succumb to more errors on the WLT than employees working container-free jobs (Hecht & Proffitt, 1995).

It is also likely that the impoverished format in which the problem was presented (i.e., on paper at a single time point) inhibited participants from utilizing their intuitive knowledge about the physical behavior of substances in the real world, leading to their erroneous
predictions. While the status of physical systems in the natural environment can be inferred
and recognized from rich (frequently updated) visual inputs, explicit tasks utilizing static
imagery convey comparatively little information. Indeed, a breadth of research in intuitive
physics has demonstrated inconsistencies between human performance on explicit and im-
plicit reasoning tasks (Kaiser et al., 1992; Kozhevnikov & Hegarty, 2001; Krist, 2000; Krist
et al., 1993; Smith et al., 2013). The pattern that emerges in these studies is that peo-
ple succumb to systematic errors when explaining a physical situation using idiosyncratic
descriptive knowledge given impoverished visual information (e.g., drawing the status of a
static situation across time) but can form accurate predictions and judgments about situa-
tions that are depicted dynamically via rich 3D visual information (Battaglia et al., 2013;
Kaiser et al., 1992; Kubricht et al., 2017; Smith et al., 2013; Ye et al., 2017). Thus, the WLT
does not appear to trigger the commonsense knowledge people have about constraints in the
physical world (e.g., early emerging sensitivity to core physical principles; see Baillargeon,
1994, 2002, 2004; Hespos et al., 2016; Spelke, 1994). Another example of discrepancy be-
tween explicit and implicit performance in intuitive physics comes from the water-pouring
problem (WPP; Schwartz & Black, 1999), a modification to the WLT which includes two
containers—one wider than the other—filled to the same height with water (see Figure 5.1B).
Participants solving the WPP must determine which container needs to be tilted farther be-
fore the water inside begins to pour out. Surprisingly, only 34% of the participants in the
study (averaged across container-type) correctly reported that a thinner container would
need to be tilted farther than a wider one. However, when instructed to complete the task
by closing their eyes and imagining the same situation, nearly all (95% of) participants ro-
tated a thinner container filled with imaginary liquid farther. These findings demonstrate
that people can reason successfully about the physical behavior of substances by mentally
simulating an imaginary eventeven if their corresponding explicit knowledge is inaccurate.
They also show that people are more likely to utilize mental simulation when those systems
are encountered in a realistic (dynamic) context instead of an ambiguous (static) one.

Taken together, the studies above show that people are capable of reasoning about the
physical status of observed situations—including those involving non-solid substances—but
fail to do so when task presentation is poor. In these cases, people appear to construct domain-specific physical theories (Cook & Breedin, 1994) which are inconsistent with their implicit expectations in the real world. Importantly, these erroneous theories can be diminished or even abolished when the problems are made less ambiguous (Kaiser et al., 1986, 1992).

The question which naturally arises is how problems can be framed to facilitate mental simulation and probabilistic inference in novel physical domains. Overall, the present study aims to address this question by focusing on judgments and predictions about physical events at critical moments: e.g., when a container will begin to spill, or where a moving substance will come to rest. Recent neural evidence suggests that people utilize an internal “physics engine” encoded within the brain’s multiple-demand system to reason about physical situations via mental simulation (Fischer et al., 2016). These events are represented “spatially” (rather than visually; see Hegarty, 2004), meaning that both observable properties (e.g., position, volume) and hidden attributes (e.g., viscosity and friction) are encoded in the inference. To facilitate spatial representation and mental simulation usage, the experiments reported herein demonstrate three key features in their design and procedure:

1. Situations are presented in a dynamic context (at least once) to guide inferences about the latent attributes and observable volumes/positions of physical entities. For example, videos demonstrating movement of each substance are shown prior to each prediction and judgment task.

2. Participants are explicitly instructed to perform mental simulations for observed situations. For example, participants are told to imagine that an event transpires before being asked about the outcome.

3. The task in each problem does not involve description or explanation of the situation across time. For example, rather than asking participants to trace out the motion trajectory of a moving object, participants are asked to catch the object or predict where the object will land at a certain time point.
In the current study, the role of dynamic context (i.e., the first rule) is particularly important because it provides visual motion cues from which substance attributes can be inferred (Kawabe et al., 2015). These attributes have been shown to influence people’s predictions about the dynamics of substances. For example, participants solving the WPP rotated imaginary containers filled with molasses—a liquid with a relatively high viscosity attribute—farther than ones filled with water (Schwartz & Black, 1999). It is therefore important to provide observers with rich visual information about substance behavior to facilitate attribute inference and subsequent mental simulation, most notably when the substance is unfamiliar.

Given that people can estimate observable and hidden properties of entities in physical scenes—and that those estimates guide predictions about future event states via mental simulation—how this is achieved requires further discussion. In the next sections, we suggest how implicit reasoning can be framed at the computational level and outline an algorithmic framework for performing this computation.

5.3 Intuitive Physics as Probabilistic Simulation

A growing body of evidence suggests that expectations about observed physical situations are formed through Bayesian inference over structured knowledge of physical principles and noisy perceptual inputs (Battaglia et al., 2013; Hamrick et al., 2016; Sanborn, 2014; Sanborn et al., 2013). This method of reasoning is surprisingly rich, evidenced by preverbal infants’ rational expectations about moving objects in complex displays (Téglás et al., 2011). The problem of inferring physical properties \( h \) can be framed as assessing the posterior probability of a candidate hypothesis \( h = H \) based on perceiving observable information \( O \), and can be computed using Bayes’ rule:

\[
P(h = H|O) = \frac{P(O|h = H)P(h = H)}{\sum_{H'} P(O|h = H')P(h = H')}. \tag{5.1}
\]

To enable the inference in Equation 5.1, a computational model needs to define the likelihood term \( P(O|h) \) for evaluating the probability of perceiving observable information given certain
physical properties, and the prior term $P(h)$ based on general knowledge of how physical properties are distributed in the world.

5.3.1 The Noisy Newton Framework

The noisy Newton framework for physical reasoning hypothesizes that inferences about complex, dynamical systems can be generated by combining noisy perceptual inputs with the principles of classical (i.e., Newtonian) mechanics $P(O|h)$, given prior beliefs about physical and perceptual variables $P(h)$ (Bates et al., 2015; Battaglia et al., 2013; Gerstenberg et al., 2015; Sanborn, 2014; Sanborn et al., 2013; Smith et al., 2013). Under this framework, the locations, motions and physical attributes of objects and substances are sampled from distributions with physical and perceptual noise and propagated forwards in time using physics-based simulation models which approximate Newtonian mechanics. The status of the situation throughout the simulation is then queried, and the outputs of the query are averaged across numerous simulations to determine the probability of the associated human judgment. The model has successfully explained a variety of human judgments across diverse physical situations, such as object collisions (Sanborn, 2014; Sanborn et al., 2013), block towers (Battaglia et al., 2013; Hamrick et al., 2016), containment situations (Liang et al., 2015), and projectile motion (Smith et al., 2013).

For simple physical events such as object collisions, the likelihood term in Equation 5.1 can be defined with four observable variables to represent head-on collision events across time (i.e., the velocity of each object before and after impact), which can be determined analytically via the principle of conservation of momentum. Physical situations in the world, however, are notably more complex and often lack computationally tractable analytic solutions. Additionally, there is often no observable evidence to gauge likelihood: e.g., when inferring future event states. For these problems, a more general framework is needed to account for a wide range of phenomena in intuitive physics.

Bates et al. (2015) extended the framework from physical scene understanding (Battaglia et al., 2013; Hamrick et al., 2016) to liquid dynamics using an intuitive fluid engine (IFE),
where future liquid states are approximated by probabilistic simulation via a Smoothed Particle Hydrodynamics method (SPH; Monaghan, 1992). The model matched human judgments about future liquid states and provided a better quantitative fit than alternative models that did not employ simulation or account for physical uncertainty.

5.3.2 Intuitive Substance Engine (ISE)

The present study developed the same general class of model as Bates et al.’s (2015) IFE, which we term the intuitive substance engine (ISE). The ISE models physical predictions and judgments by simulating substance states forwards in time and querying perceptual variables at critical time steps. Substance states are represented by the perceptual and physical variables that define their physical behavior, such as the position of each substance element and the physical attributes (e.g., viscosity, density, pressure, etc.) which govern how those positions change over time. The state of a substance at time step \( t \) is denoted by \( S_t \), where \( t = 0, 1, \ldots, T \). Given an initial ground-truth substance state, \( \bar{S}_0 \) (i.e., the true values for each perceptual and physical variable prior to movement), the ISE first forms an observed state distribution, \( S_0 \), reflecting noisy perception and prior beliefs about underlying variables: \( P(S_0|\bar{S}_0) \). This distribution is then simulated forwards in time using a physics-based simulation method, and the state distributions over time are queried to form predicted response distributions. A graphical depiction of the ISE modules is shown in Figure 5.2.

Although both our ISE and Bates et al.’s (2015) IFE are formed under the same computational framework, each employs different simulation methods. The simulation of incompressible flows through numerical evaluation of physical equations has become one of the most significant topics in computer graphics and mechanical engineering. The velocity field of simulated substances is determined according to the constraints specified in the Navier-Stokes equations. These partial differential equations place constraints on key physical properties (i.e., momentum and compressibility) which are quantified by underlying variables, such as localized substance velocity, density, pressure, and viscosity (see Appendix C). To numerically solve these equations, we adopt the material point method (MPM; Jiang et al., 2015;
Figure 5.2: ISE Modules. The three core modules of the ISE model are shown. (i) Input variables are separated into perceptual (observable) and physical (hidden) variables. Prior distributions are placed on the listed variables; other perceptual and physical variables are also passed to the simulator (e.g., flow velocity, density, etc.) but only their ground-truth values were used. (ii) Substance states follow noisy distributions due to uncertain prior expectations about underlying variables. The ISE model uses the sampling approach with MPM simulation to approximate the substance state distribution at each time step: \( P(S_{0:T}) \). (3) This distribution is queried based on the question asked to participants in each experiment, and queries are aggregated across \( n = 1, ..., N \) samples to form predicted response distributions.
Zhu & Bridson, 2005), which has become standard in physics-based simulation calculations due to its accuracy, stability and efficiency.

The MPM produces physically accurate and visually realistic simulations of the dynamics of liquid (Jiang, Schroeder, Teran, Stomakhin, & Selle, 2016) and sand (Klár et al., 2016), in addition to general continuum materials such as stiff elastic objects (Jiang et al., 2016). Unlike Bates et al.’s SPH method which purely relies on particles to discretize the computational domain, the MPM uses both particles and a background Eulerian grid. The Navier-Stokes equations are solved on the grid, allowing for: (i) accurate derivative calculations; (ii) well-defined free surface and solid boundary conditions; and (iii) an accurate first-order approximation of physical reality. MPM also circumvents common artifacts of SPH: e.g., underestimated density near free surfaces and weakly compressible artifacts. In fact, the requirement for incompressibility is crucial in the liquid-pouring problem studied in the present study. We choose not to use SPH because it does not guarantee a divergence-free velocity field unless additional computational components are included. MPM, however, maintains the benefits of particle-based methods due to its hybrid particle/grid nature. The presence of particles in the current model serves to facilitate visualization and the tracking of material properties. Besides modeling liquids, the state-of-the-art physics-based simulation methods have also provided realistic cues for modeling complex tool and tool-uses (Zhu, Zhao, & Zhu, 2015), generic containers (Liang et al., 2015), and soft human body dynamics (Zhu, Jiang, Zhao, Terzopoulos, & Zhu, 2016). Appendix C presents a mathematical overview of our MPM simulator, which provides a unified, particle-based simulation framework that handles rigid balls, liquid, and sand with essentially the same numerical algorithm, albeit with appropriately differing material parameters. The MPM method is physically accurate, numerically stable, and computationally efficient, enabling us to synthesize a large set of stimuli in a short amount of time by simply varying material parameters and the locations of the initial objects and colliding geometries. Running each simulation in the same framework for the purposes of the present study also enables fair comparisons among the three types of substances, since we avoid potential inconsistencies in the numerical accuracies of multiple simulators specialized to each material.
5.3.3 Perceptual Noise and Uncertainty

Fluid simulation with physical dynamics provides deterministic fluid movements if the ground-truth values of substance attributes, position, and volume are known. Hence, the decisions directly derived from the MPM simulator are binary judgments, which implies that physical simulation with high precision cannot explain humans’ probabilistic judgments in intuitive physics tasks. Inspired by the approach of Bates et al. Bates et al., 2015 and the noisy Newton model (e.g., Sanborn et al., 2013) we combine the MPM simulator with noisy input variables (i.e., position, volume, viscosity, friction angle, and restitution) to form the ISE model, thereby accounting for physical uncertainty and the influence of perceptual and physical variables on our prediction and judgment tasks. Additional details on the ISE are provided in Appendix C, and the distributions used to generate noise—along with their parameters—are provided in the ISE Model Details section in each experiment.

It is important to note that our ISE (employing MPM simulation) is roughly equivalent to Bates et al.’s (2015) IFE (employing SPH simulation) in that both models apply the noisy Newton framework to substance dynamics. Indeed, SPH is a viable method for simulating the dynamics of both granular materials and liquids, although MPM provides a more efficient and accurate means of doing so. We do not envision that the predictions of the two methods would differ substantially from one another when applied to a given set of stimuli.

The present study aims to determine whether an approximated simulation model coupled with noisy input variables can account for human predictions and judgments in three novel situations for non-solid substances. In the first situation (Experiment 1), participants reasoned about the relative pouring angle of two containers filled with liquids differing in their volume and viscosity. The first experiment was inspired by previous empirical findings in water-pouring tasks—namely that people take physical attributes such as viscosity into account when making liquid-related judgments (Schwartz & Black, 1999). Bates et al. (2015) also found that their participants’ judgments were sensitive to latent attributes of the observed liquid (e.g., stickiness and viscosity). In the second experiment, participants reasoned about the dynamics of sand—a granular substance less ubiquitous in daily life than
viscous liquids. In the third and final experiment, participants reasoned about the dynamics of liquid, sand, and rigid balls in a task adapted from Bates et al.’s (2015) experiments.

Our study—which uses a modification of Schwartz and Black’s (1999) water-pouring problem and Bates et al.’s (2015) basin prediction problem—utilizes advanced techniques in computer graphics to test the hypothesis that humans can form predictions and judgments about non-solid substances when provided with animated demonstrations of flow behavior which guide the inference of unobservable attributes. We hypothesize that these inferred attributes inform mental simulations and enhance performance in subsequent reasoning tasks. The human capacity for predicting and judging the dynamics of rigid bodies has been demonstrated across a variety of situations in recent years (Battaglia et al., 2013; Gerstenberg et al., 2015; Hamrick et al., 2016; Sanborn, 2014; Sanborn et al., 2013; Smith et al., 2013; Téglás et al., 2011). However, the physical equations governing substance dynamics are notably more complex than those governing the behavior of rigid bodies. For example, in a situation comprised of stacked blocks or colliding balls, it is clear which entities need to be represented to form meaningful expectations. If the number of entities is small enough, underlying physical constraints can even be emulated by deep networks which learn how those entities interact with one another over time (Battaglia et al., 2016; Grzeszczuk et al., 1998). However, our MPM simulator approximates substance volumes as a collection of numerous particles with high spatial resolution, which clearly exceeds human capabilities. We further test the hypothesis that people represent substances as discrete collections of rigid balls—on the order of ten to twenty—and form predictions by applying emulated principles of rigid-body mechanics to spatially represented variables. We term this the ball approximation model (BAM) and test it alongside our ISE in the General Discussion section (Section 5.7).

Across three experiments, the current study aims to provide the following contributions: (i) provide evidence that people’s predictions and judgments about the dynamics of substances are consistent with a computational model utilizing approximated Newtonian principles and a noisy perception model, (ii) demonstrate that intuitive beliefs about substance dynamics cannot be captured by simple heuristics or data-driven (non-simulation) models, (iii) show that when observing visually identical stimuli, peoples predictions about the dy-
namics of liquid, sand, and rigid objects differ from one another, and (iv) discern whether a simulation method which approximates substances as a collection of rigid balls can capture trends in human performance in lieu of advanced numerical simulation methods.

5.4 Experiment 1: Reasoning about the Relative Pouring Angle of Liquid-Filled Containers

Our primary motivation for Experiment 1 is to demonstrate that people can utilize mental simulation to form judgments about the dynamics of liquids which vary in their viscosity attribute. The experimental task is designed to conform with the three features outlined in Section 5.2 to facilitate spatial representation and subsequent simulation. To quantify the extent that people employ inferred attributes of non-solid substances when reasoning about novel situations, we utilized a recent development in graphical substance simulation (Bridson, 2008; Jiang et al., 2015) to capture the dynamic behavior of non-solid materials in vivid animations. Previous work has shown that realistic animations can facilitate representation of dynamic physical situations (Tversky & Morrison, 2002). Furthermore, recent research on human visual recognition indicates that latent attributes of liquids (e.g., viscosity) are primarily perceived from visual motion cues (Kawabe et al., 2015). Therefore, displaying realistic substance behavior can provide the input of key physical attributes that enable mental simulation.

5.4.1 Method

Participants. A total of 152 participants were recruited from the Department of Psychology subject pool at the University of California, Los Angeles, and were compensated with course credit.

Materials and Procedure. Prior to the reasoning task, participants viewed animated demonstrations of the movement of a moderately viscous liquid in two situations. The liquid used in the demonstrations was colored orange and was not observed in the judgment task. In
the first (flow) demonstration, the orange liquid pours over two torus-shaped obstructions in a video looped three times and lasting for 11.5 seconds. The flow demonstration videos were presented to provide visual motion cues to inform participants’ inferred viscosity values. Following the flow demonstration, participants viewed a video of a cylindrical container filled with the same orange liquid tilting at a constant angular rate \((\omega = 22^\circ \cdot \text{sec}^{-1})\) from the upright orientation of the container and moving towards the horizontal. The tilting demonstration video was looped three times for a duration of 14.7 seconds.

Following the demonstration videos, two new liquids were introduced, one with low viscosity \((L_{visc};\text{similar to water})\) and one with high viscosity \((H_{visc};\text{similar to a thin syrup})\). The \(L_{visc}\) and \(H_{visc}\) liquids were colored either red or green, and the color was counterbalanced across participants. As shown in the top panel of Figure 5.3, participants viewed a flow demonstration video of both the \(H_{visc}\) and \(L_{visc}\) liquids (looped three times) for a duration of 11.5 seconds before each judgment trial. The two flow videos were presented side by side for comparison, and the relative position of each liquid was counterbalanced across participants. The \(L_{visc}\) and \(H_{visc}\) liquids were selected to readily distinguish each one based on their perceived viscosities, which were inferred from visual motion cues in the flow demonstration videos; see (Kawabe et al., 2015).

In the subsequent reasoning task, participants viewed a static image of two containers side by side filled with the \(L_{visc}\) and \(H_{visc}\) liquids (see bottom panel of Figure 5.3). Participants were instructed to assume that each container was tilted simultaneously in the same way as observed earlier for the orange liquid in the tilting demonstration. They were informed that both containers were tilted at the same rate, and were provided with the quantity of liquid in each container. Participants were then asked to report which container would need to be tilted with a larger angle before the liquid inside begins to pour out and received no feedback following completion of each trial. The experiment manipulated the volume of the \(L_{visc}\) and \(H_{visc}\) liquids \((V_L\text{ and }V_H, \text{respectively})\) in each container across the values 20%, 40%, 60%.

---

1 The flow demonstration video can be viewed at https://vimeo.com/264205657
2 The tilting demonstration video can be viewed at https://vimeo.com/264206162
3 The stimulus videos can be viewed at https://vimeo.com/264204243
Figure 5.3: Flow Demonstration Video and Judgment Trial (Experiment 1). (Top) Sample frames from the *Hvisc* (red) and *Lvisc* (green) flow demonstration video. (Bottom) Illustration of tilt judgment trial, where the proportion of each container filled with the *Hvisc* and *Lvisc* liquid is 40% and 60%, respectively.

and 80%, representing the proportion of the container filled. Hence, the experiment consisted of 16 trials presented in a randomized order, including all possible volume pairs between the *Lvisc* and *Hvisc* liquids. The experiment lasted approximately 10 minutes.

5.4.2 Results

**Human Results.** The proportion of participants choosing the *Hvisc* liquid container as pouring last for each judgment trial is shown in Figure 5.4. To assess the relationship between *Hvisc* liquid volume and human judgments, a repeated-measures ANOVA was conducted across two within-subjects factors: i.e., *Lvisc* and *Hvisc* liquid volume with four levels each. The two-way interaction between *Lvisc* and *Hvisc* liquid volume was significant—\(\eta^2_{p} = 0.74, F(9, 143) = 45.12, p < .001\)—indicating that the effect of *Hvisc* liquid volume on *Hvisc* response proportion varied according to the quantity of *Lvisc* liquid in the alternative
Figure 5.4: Human Results (Experiment 1). Human Hvisc response proportions for all experimental conditions. The volume of the Hvisc fluid ($V_H$) is plotted on the horizontal axis, and separate lines indicate the four possible volumes for the Lvisc fluid ($V_L$).

can. Since the two liquid volume variables interacted, data were separated into each Lvisc liquid volume condition and the main effect of Hvisc liquid volume was examined. Results indicate that Hvisc liquid volume was least influential on participants’ judgments when the Lvisc container was 80% filled: $\eta^2_p = 0.05$, $F(3, 149) = 2.36$, $p = .07$. Hvisc liquid volume had a comparatively greater effect for smaller Lvisc volumes: $V_L = 60\%$, $\eta^2_p = 0.59$, $F(3, 149) = 70.77$; $V_L = 40\%$, $\eta^2_p = 0.74$, $F(3, 149) = 143.12$; $V_L = 20\%$, $\eta^2_p = 0.68$, $F(3, 149) = 107.22$. In other words, when the Lvisc liquid container had greatest volume, no matter how much Hvisc liquid was in the alternative container, participants consistently reported that the Hvisc container would pour last. However, participants were increasingly hesitant to report that the Hvisc liquid container would pour last when the Lvisc liquid container was filled to lesser volumes. These results indicate that participants attended to both liquid volume and viscosity when forming their relative pour angle judgments.

Can Heuristic-Based Reasoning Account for Human Performance? Next, we examined whether people rely on heuristic-based reasoning to make their judgments. One
candidate heuristic is that given two containers filled with different volumes of each liquid, the container with lesser liquid volume requires a greater rotation before beginning to pour. While participants adhered to this rule for trials where \( V_H < V_L \), their judgments for each of the \( V_L < V_H \) trials did not accord to the same heuristic. For example, in trials where volume difference was least salient (i.e., \( V_H = 40\%, 60\%, \) and \( 80\% \); and \( V_L = 20\%, 40\%, \) and \( 60\% \), respectively), the lesser-volume heuristic predicts \( L_{visc} \) liquid responses. However, \( H_{visc} \) response proportions for those trials were significantly greater than zero: \( t(151) = 9.92, 8.86, 8.10 \); Cohen’s \( d = .80, .72, .66 \), respectively. A second heuristic is to always choose the \( H_{visc} \) fluid as requiring a greater rotation since it moves slower than the \( L_{visc} \) fluid. The above three cases also disagreed with this heuristic since \( H_{visc} \) response proportions were significantly less than one: \( t(151) = 15.22, 17.04, 18.65 \); Cohen’s \( d = 1.23, 1.38, 1.51 \), respectively. In summary, response proportions in the specified trials reveal that participants attended to latent fluid attributes (e.g., viscosity) and volume difference when making their tilt angle judgments.

**ISE Model Details.** The input variables for the ISE in Experiment 1 were liquid volume with perceptual uncertainty and liquid viscosity with physical uncertainty. Given the ground-truth value of volume and a biased value of viscosity for each liquid \( (V_L^{(GT)}, V_H^{(GT)}, \mu_L^{(B)}, \mu_H^{(B)}) \), \( N = 10,000 \) noisy samples \( \{(V_l^{(i)}, V_h^{(i)}, \mu_l^{(i)}, \mu_h^{(i)}), i = 1, ..., N\} \) were generated and passed to the MPM simulator. The simulator propagated each sampled situation forwards in time and determined when each container’s contents began to spill over the rim of the cylinder. The container which required a longer duration to spill was chosen as the predicted response for each sample. By aggregating predictions across the 10,000 samples, the ISE outputs a predicted response distribution for each trial.

To model perceptual and physical uncertainty in participants’ mental simulations, the ISE sampled liquid volumes and viscosities from noisy distributions reflecting imperfect volume estimation and viscosity inference via the visual system. Gaussian noise (0 mean, \( \sigma_V^2 \) variance) was added to the ground-truth volume—\( V_L^{(GT)} \) and \( V_H^{(GT)} \)—in each trial. Gaussian noise (0 mean, \( \sigma_\mu^2 \) variance) was also added to a biased viscosity value for each liquid—\( \mu_L^{(B)} \)
and $\mu_B^{(B)}$. These parameters were free to vary in the ISE model and were chosen instead of the ground-truth viscosity values to account for biased inferences. For example, if people have a prior belief that liquids should behave like water, their inferred viscosities should be negatively biased for the $H_{visc}$ liquid. Viscosity uncertainty was added in logarithmic space (Sanborn et al., 2013): $\mu_i = \log^{-1}(\log(\mu^{(B)}_i) + \varepsilon)$, where $\mu^{(B)}_i$ is the biased viscosity value and $\varepsilon$ represents Gaussian noise with 0 mean and $\sigma_\mu^2$ variance. Logarithmic noise corresponds with increasing uncertainty for larger variable values. The results reported herein used the following model parameters: $\mu_L^{(B)} = 2.5$, $\mu_H^{(B)} = 110$, $\sigma_V = 0.15$, and $\sigma_\mu = 0.15$.

Non-Simulation Model Details. To examine whether substance simulation is necessary to account for how humans reason about liquid dynamics, we compare the ISE model with two statistical learning methods: the generalized linear model (GLM; McCullagh, 1984) and eXtreme Gradient Boosting (XGBoost; Chen & Guestrin, 2016). These models are purely data-driven and do not involve any explicit knowledge of physical laws or physical simulation. The selected features for these models include: (i) the volume of liquid in each container; and (ii) the viscosity value of the $L_{visc}$ and $H_{visc}$ liquids.

To predict the human judgment for the $i^\ast$th trial $J_{i^\ast}$, both non-simulation models were trained with the remaining 15 trials $\{J_i, i = 1, 2, ..., 16, i \neq i^\ast\}$ and tested with the $i^\ast$th trial. The trained GLM model is directly applied to the test case to predict which container will need to be tilted to a larger angle before the liquid inside begins to pour out. Since the XGBoost model is only capable of making a (direct) discriminative classification (i.e., +1 indicating selection of the left container and -1 indicating selection of the right container), we introduced perceptual noise (the same method for the ISE) to each test trial to model perceptual and physical uncertainty. For each test trial, a set with 10,000 samples was generated. The trained XGBoost model is applied to classify the labels (+1 or -1) in each sample, which are then aggregated to form the predicted response distribution for each test trial.

Model Comparisons. We first compared how well different computational models account for human performance for the 16 trials. Figure 5.5 depicts results from the ISE, GLM, and
XGBoost models with perceptual noise. Human judgments and model predictions were highly correlated: $r(14) = 0.995, 0.95, \text{ and } 0.93$, respectively. Root-mean-squared deviation (RMSD) between human judgments and the models’ predictions were 0.089, 0.13, and 0.12, respectively. Unlike the purely data-driven models (i.e., the GLM and XGBoost models), the simulation-based ISE model utilizes material properties (e.g., viscosity) and perceptual features (e.g., volume) as variables in a generative physical model and provides better approximations to human judgments in the viscous liquid-pouring task. These results again support the role of simulation as a potential mental model that supports human inference in physical reasoning tasks.

Next, we examined whether the ISE model captures human performance on the three trials where the heuristic rule outlined earlier provides incorrect predictions; i.e., those trials where $V_H = 40\%, 60\%, \text{ and } 80\%$ and $V_L = 20\%, 40\%, \text{ and } 60\%$, respectively. Here, the ground-truth model predicts that the $H_{visc}$ liquid will require a greater angle of rotation before beginning to pour, while the heuristic rule discussed earlier suggests the opposite. $H_{visc}$ response proportions for these trials were 0.39, 0.34, and 0.30, respectively, and the ISE model returned consistent predictions of 0.51, 0.35, and 0.24 (95% CIs [0.48, 0.55], [0.33, 0.41], and [0.16, 0.26]). Alternatively, the GLM model predicted response proportions of 0.48, 0.45, and 0.49 (95% CIs [0.45, 0.51], [0.42, 0.49], and [0.46, 0.51]), and the XGBoost model predicted response proportions of 0.49, 0.47, and 0.47 (95% CIs [0.45, 0.51], [0.45, 0.49], and [0.45, 0.51]). These results indicate that the non-simulation models consistently predicted response proportions biased towards the ground-truth model and away from the lesser-volume heuristic prediction. Since the number of free parameters varied between the ISE and non-simulation models, we chose to compare Bayesian information criterion (BIC) values for each model. This measure quantifies model accuracy and is penalized for model complexity; lower values indicate a superior model fit. BIC values for the ISE, GLM and XGBoost models are -730.3709, -612.4926, and -623.7055, respectively. In summary, our ISE model captured human performance on the specified trials and outperformed competing models after adjusting for model complexity.
Figure 5.5: **Model Results (Experiment 1).** Comparison of results between the three prediction models: (Left) ISE, (Middle) GLM, and (Right) XGBoost with perceptual noise. (Top) Horizontal axes indicate $HV_{\text{visc}}$ liquid volume; vertical axes indicate the predicted proportion of $HV_{\text{visc}}$ liquid responses associated with a greater rotation angle. (Bottom) Model predictions vs. human judgments. Vertical lines indicate the CIs of human judgments; horizontal lines indicate the CIs of model predictions. The ISE simulation model outperforms competing non-simulation models.

### 5.5 Experiment 2: Reasoning about Granular Materials

Results from Experiment 1 indicate that people can simulate the dynamics of non-solid substances (i.e., liquids) which vary in their viscosity attribute. However, it remains unclear whether this proficiency extends to substances that are less ubiquitous in daily life than liquid—specifically granular materials such as sand. The second experiment was designed to determine whether humans can predict the resting geometry of a volume of sand after it is poured from a funnel onto a surface, and whether dynamic visualizations of the pouring behavior facilitate mental simulation of sand-surface interactions.
5.5.1 Method

Participants. A total of 108 undergraduate students (81 females; mean age = 20.2 years) were recruited from the University of California, Los Angeles (UCLA), Department of Psychology subject pool and were compensated with course credit.

Materials and Procedure. Participants first viewed a simulated demonstration video of sand falling from a funnel suspended 10 cm above a level surface. The pouring event was viewed three times from a zoomed-out perspective (Figure 5.6A) and then a zoomed-in perspective (Figure 5.6B). The duration of the video was 35 sec\(^4\). After viewing the demonstration video, participants were shown a sand-filled funnel suspended 1/2, 1, 2, and 4 cm above the surface in a randomized order.

\(^4\)The demonstration video can be viewed at https://vimeo.com/267150402
Forty-three participants were assigned to the Static Condition and viewed a static image (zoomed-out) in which the funnel was positioned at each of the indicated heights. Sixty-five participants were assigned to the Dynamic Condition and viewed a video (zoomed in and out; looped three times; 35 sec duration) of sand pouring from a funnel that was positioned at the aforementioned heights above the surface. In the Dynamic Condition, the region of the surface where the sand fell was occluded by a gray rectangle.

After viewing each situation, participants were asked to indicate which of four sand piles would result from the sand pouring from the funnel at the indicated height (Figure 5.6C). For each trial, the stimulus images (for the Static Condition) and final video frames (Dynamic Condition) remained on the screen until a response was made. The pile choices were shown from the zoomed-in perspective and represented the ground-truth resting geometries resulting from each situation: i.e., Piles 1, 2, 3, and 4 correspond with the piles resulting from sand pouring from funnels suspended 1/2, 1, 2, and 4 cm above the surface, respectively. The experiment consisted of 4 trials presented in a randomized order.

5.5.2 Results

**Human Results.** Participants’ pile choices varied across different heights, $\chi^2(9) = 176.54$; Cramer’s $V = 0.74$, indicating that funnel height influenced their predictions about the resting geometry of falling sand. The impact of funnel height on the judgment of the resulting sand pile revealed the involvement of physical reasoning in this task. In addition, the proportion of participants choosing each sand pile did not differ between the Dynamic and Static Conditions: $\chi^2(3) = 2.21, 2.34, 2.41,$ and $1.13$; Cramer’s $V = 0.14, 0.15, 0.15, 0.10$ for funnel heights of 1/2, 1, 2, and 4 cm, respectively. These results suggest that dynamic visualizations of sand pouring from the funnel in each situation did not provide sufficient information to further facilitate physical reasoning in the task.

As shown in Figure 5.7, participants’ pile choices shifted towards higher-numbered, flatter piles as funnel height increased. These results indicate that participants’ predictions were sensitive to funnel height, but inconsistent with ground-truth resting states. In the following
Figure 5.7: **Human and Model Results (Experiment 2).** Model prediction results compared to human judgments. (Upper) Static Condition and (Lower) Dynamic Condition. Each bar (1, 2, 3, and 4) corresponds with testing trials with funnel height 1/2, 1, 2, and 4 cm, respectively.

sections, predictions from the three computational models outlined earlier (ISE, GLM, and XGBoost) are reported and compared to human performance to determine whether the noisy Newton framework can account for participants’ deviations from ground-truth judgments.

**ISE Model Details.** The input variables for the ISE in Experiment 3 were funnel height (i.e., initial sand height) with perceptual uncertainty and sand friction angle with mental simulation uncertainty. Given the ground-truth values of initial funnel height and friction angle \( (H^{(GT)}, \theta^{(GT)}) \), \( N = 10,000 \) noisy samples \( \{(H^{(i)}, \theta^{(i)}), i = 1, ..., N\} \) were generated and passed to the MPM simulator, which returned the final height of the sand pile for each sample. Instead of choosing from 4 piles (i.e., the task presented to the participants), the MPM simulator compares the estimated height of the final sand pile, formally \( D(H^{(i)}, \theta^{(i)}) = H_p \in \mathbb{R} > 0 \), with the heights of the 4 pile options given to participants. The pile option
with the minimum height difference was chosen as the predicted judgment for each sample.
Finally, by aggregating predictions across the 10,000 samples, the ISE outputs a predicted
response distribution for each trial.

To model physical uncertainty in participants’ mental simulations, the ISE sampled funnel
heights and friction angles from noisy distributions. Gaussian noise (0 mean, \( \sigma_H^2 \) variance) was added to the ground-truth funnel height in each situation. Gaussian noise was also added to the ground-truth friction angle \( \theta^{(GT)} \), but in logarithmic space (Sanborn et al., 2013):

\[
\theta^{(i)} = f^{-1}(f(\theta^{(GT)}) + \varepsilon),
\]

where \( \theta^{(GT)} \) is the ground truth value of the initial sand height, \( f(\theta^{(GT)}) = \log(\omega|\theta^{(GT)}|) \), and \( \varepsilon \) represents Gaussian noise with 0 mean and \( \sigma_\varepsilon^2 \) variance. The results reported herein used the following model parameters: \( \sigma_H = 0.11 \), \( \sigma_\varepsilon = 0.65 \), and \( \omega = 0.85 \).

Non-Simulation Model Details. To examine the crucial role of mental simulation in physical judgment, two non-simulation models, GLM and XGBoost, were used as baseline models. The two models were trained on the three piles and tested on the remaining \( i^* \)th pile \((i = 1, 2, 3, 4, i \neq i^*)\). During training, 10,000 samples were drawn for each remaining pile (30,000 samples total) and passed to the MPM simulator. Samples were generated using the sampling method described in the previous section. After training on the 30,000 samples, both non-simulation models were tested on another 10,000 samples generated from noisy input based on the configuration of Pile \( i^* \). The final distribution was formed by aggregating the predictions across the 10,000 samples.

Model Comparisons. Figure 5.7 depicts the predictions of the ISE, XGBoost, and GLM models compared to human judgments. All four models achieved high correlations with human performance, Static: \( r(12) = 0.91, 0.86, \text{ and } 0.77 \); Dynamic: \( r(12) = 0.88, 0.87, \text{ and } 0.73 \) for ISE, XGBoost, and GLM, respectively. Human performance was much less correlated with ground-truth predictions, Static: \( r(12) = 0.17 \); Dynamic: \( r(12) = 0.19 \). The ISE model predictions were more correlated with the human data than the competing data-driven model predictions in the Static condition but were only slightly more correlated than XGBoost predictions in the Dynamic condition. Hence, this paper uses the root-
Table 5.1: **Model Performance Measures (Experiment 2).** RMSD values, along with 95% confidence intervals on the probability of choosing each pile are provided for the ISE, XGBoost, and GLM models in the (Left) Static and (Right) Dynamic Conditions. The number of free parameters and corresponding BIC measures are also included for each model. Note that lower BIC values indicate superior model performance.

<table>
<thead>
<tr>
<th></th>
<th>Static</th>
<th>Dynamic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ISE</td>
<td>XGBoost</td>
</tr>
<tr>
<td>RMSD</td>
<td>0.101</td>
<td>0.183</td>
</tr>
<tr>
<td>CIs (95%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pile 1</td>
<td>[0.73, 0.81]</td>
<td>[0.98, 1.00]</td>
</tr>
<tr>
<td>Pile 2</td>
<td>[0.09, 0.15]</td>
<td>[0.01, 0.05]</td>
</tr>
<tr>
<td>Pile 3</td>
<td>[0.01, 0.05]</td>
<td>[0.15, 0.21]</td>
</tr>
<tr>
<td>Pile 3</td>
<td>[0.27, 0.32]</td>
<td>[0.22, 0.26]</td>
</tr>
<tr>
<td>Params.</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>BIC</td>
<td>-476.48</td>
<td>-348.10</td>
</tr>
</tbody>
</table>

Mean-square deviation (RMSD) between human responses and model results to compare the model fits. We found that RMSD between human responses and ISE predictions for the four judgment trials was less than that between ground-truth predictions in both the Static and Dynamic conditions (see Table 5.1). We also examined model performance using the Bayesian information criterion (BIC) to account for the different number of free parameters in each model. We found that the ISE provides a better fit to the human data than the ground-truth and data-driven models in both conditions. For ISE, XGBoost, and GLM models, Static BIC = -481.16, -348.10, -339.06, and Dynamic BIC = -462.73, -379.82, -394.30, respectively. The model with the lowest BIC value is preferred.
5.6 Experiment 3: Reasoning about Complex Interactions Between Sand and Rigid Obstacles

Our results from Experiment 2 indicate that people can predict the resting geometry of sand poured from a funnel, even though they may not have rich experience interacting with granular substances in daily-life. Experiment 3 was conducted to determine (i) whether humans can reason about complex interactions between sand and rigid obstacles; and (ii) whether their predictions about the resting state of sand in novel situations differ from predictions about other more familiar substances, such as liquid and rigid balls.

5.6.1 Method

Participants. A total of 90 undergraduate students (66 females; mean age 20.9), were recruited from the UCLA Department of Psychology subject pool, and were compensated with course credit.

Materials and Procedure. The procedure in Experiment 3 was similar to the design in Bates et al.’s (2015) experiment: i.e., participants viewed a quantity of a substance suspended in the air above obstacles and were asked to predict the proportion that would fall into two basins separated by a vertical divider below (see Figure 5.8). The present experiment differed from previous work in that participants reasoned about the resting state of one of three different substances: liquid, sand, or sets of rigid balls. Also, whereas the previous study used polygonal obstacles, those in the present study were circles varying in size. Depth information was also not present in the rendered situations⁵.

Situations were generated by sampling between 2 and 5 obstacle locations from a uniform distribution bounded by the width and height of the chamber. The diameter, d, of each obstacle was sampled from a uniform distribution: \( d \sim U(0.15, 0.85) \). The center points for each set of obstacles were generated by uniformly sampling the entire width and height of the chamber. If the generated obstacles were placed outside of the boundary, the configuration

⁵The stimulus videos can be viewed at https://vimeo.com/216585992
Figure 5.8: Judgment Trial for Each Substance (Experiment 3). Initial (top) and final (bottom) state of liquid (left), sand (middle), and a set of rigid balls (right) for a testing trial in Experiment 3 with 4 obstacles. The number of obstacles varied between 2 and 5 in the testing trials. Percentages indicate the amount of each substance that fell into the left and right basins. Only the initial state of each substance was shown in the testing trials.

was rejected and re-sampled. Our MPM simulator was used to determine the ground-truth proportion of each substance in the left and right basins for each of the generated situations. For each substance, forty testing trials (10 trials with 2, 3, 4, and 5 obstacles) were chosen from the generated set such that the ground-truth proportion of liquid in the left basin was approximately uniform across trials: $L \sim U(0,1)$. Importantly, the initial configuration of the testing trials (i.e., the positions of the substance and obstacles) were the same for each substance.

Participants were randomly assigned to either the liquid, sand, or rigid balls condition. Thirty participants were assigned to each condition in a between-subjects experimental design. Prior to the testing trials, participants completed five practice trials with two obstacles in each situation in a randomized order. After answering (i) which basin the majority of the substance would fall into and (ii) the expected proportion that would fall into the indicated basin, participants viewed a video (13 second duration) of the situation unfolding and were told the resulting proportion in the ground-truth simulation. The videos were shown to
provide participants with visual information to infer each substances respective attribute: i.e., viscosity, friction angle, or restitution. After completing the practice trials, participants completed 40 testing trials in a randomized order by answering the same two questions in each trial. No feedback was given following the completion of each testing trial.

5.6.2 Results

**Human Results.** Participants’ predicted (left-basin) proportions in the testing trials were strongly correlated with ground-truth predictions in the liquid, sand, and rigid balls conditions, $r(38) = 0.86, 0.82, \text{ and } 0.88; \text{RMSD} = 0.145, 0.170, 0.186,$ respectively. The deviation for each trial was calculated by subtracting the ground-truth proportion from each participant’s proportion response. The deviation differed significantly between the three substance conditions, $F(2) = 3.64, p = 0.03,$ indicating that the difference between human predictions and the ground-truth status varied according to the substance type. To determine whether participants’ response proportions differed between substances, a random factor ANOVA was conducted for a chosen set of trials. The chosen set excluded those trials where the majority of each substance fell into the same basin (left or right) according to the ground-truth simulation. We found that the response proportions showed significant differences depending on substance type, $F(2) = 8.43, p < 0.01.$ The next section examines whether the ISE and two non-simulation models can capture differences in human performance between the three substances.

**ISE Model Details.** In Experiment 3, the observable input variables for our ISE for each substance were (1) the initial, horizontal location of the substance; and (2) the locations of the circular obstacles in each situation. The latent substance attributes accepted by the engine were viscosity, friction angle, and restitution coefficient for liquid, sand, and the rigid balls, respectively. Given ground-truth values of substance position, obstacle position (2-5 obstacles), and substance attributes ($L_{S}^{(GT)}, L_{O}^{(GT)}; \mu^{(GT)}, \theta^{(GT)}, \text{ or } e^{(GT)})$, $N = 2000$ samples (40 situations $\times$ 50 noisy samples) were generated for each substance ($\{L_{S}^{(i)}, L_{O}^{(i)}; \mu^{(i)}, \theta^{(i)}, \text{ or } e^{(i)}\}, i = 1, N$) and passed to the MPM simulator. The simulator returned the final proportion of each substance that fell into the left basin.
To model uncertainty in the perceptual and physical inputs, Gaussian noise was added to each substances (ground-truth) horizontal position (0 mean, $\sigma_S^2$ variance) and the obstacles’ (ground-truth) positions in 2D space (0 mean, $\sigma_O^2$ variance). Logarithmic Gaussian noise was added to each substances ground-truth attribute value via the logarithmic transformation specified in Experiment 1. The same noise parameter was used for each substance attribute. The results reported here utilized the following model parameters: $\sigma_S = 0.59$, $\sigma_O = 0.63$, $\sigma_\varepsilon = 0.5$, $\omega = 0.8$.

**Non-Simulation Model Details.** Like the previous experiments, both GLM and XGBoost were trained and tested on the judgment trials. The training data were randomly generated situations (see above scene generation procedure) with basin proportions calculated using resting state output from our MPM simulator. Input features were the collection of both the observable input variables and latent substance attributes used in the ISE prediction. In total, 6000 samples were used for training.

**Model Comparisons.** Figure 5.9 depicts the comparison between human and model basin predictions from the ground-truth (GT), ISE, GLM, and XGBoost models, and Table 1 depicts the root-mean-square deviation (RMSD) of each model’s predictions from human ones. The human data were highly consistent with ISE predictions, $r(38) = 0.93, 0.93, 0.93$; RMSD = 0.081, 0.080, 0.102 for liquid, sand, and rigid balls, respectively. The ISE model predictions deviated from the human data to a lesser degree than the GT model predictions—$r(38) = 0.92, 0.92, 0.93$; RMSD = 0.082, 0.085, 0.104 for liquid, sand, and rigid balls, respectively—indicating a superior account of human predictions across a range of substances. In comparison, GLM and XGBoost predictions were less consistent with human predictions, GLM: $r(38) = 0.67, 0.74, 0.71$, RMSD = 1.382, 1.422, 2.067; XGBoost: $r(38) = 0.77, 0.78, 0.64$, RMSD = 0.098, 0.120, 0.193 for liquid, sand, and rigid balls, respectively. As in the previous experiment, we compared our models’ BIC measures in each condition to account for the number of free parameters in each model. We found that the BIC values for the ground-truth, GLM, and XGBoost models—GLM: BIC = 76.23, 81.49, 148.70; XGBoost: BIC = -404.60, -366.65, -283.55 for liquid, sand, and rigid balls, respectively—
were consistently greater than the values for the ISE model—BIC = -426.44, -416.60, -383.47 for liquid, sand, and rigid balls, respectively—further reinforcing the superior performance of our simulation-based ISE model.

It is worth noting that our ISE achieved consistent performance across all three substances, whereas GLM and XGBoost were less capable of predicting human judgments about rigid balls and liquid. In addition, our ISE used only one third of the training samples that XGBoost and GLM needed, demonstrating that a generative physical model with noisy perceptual inputs is capable of learning with a smaller number of samples than data-driven methods.

5.7 General Discussion

Results from the experiments reported herein provide converging evidence that humans can predict outcomes of novel physical situations involving non-solid substances by propagating approximate spatial representations forward in time using mental simulation. This stands in
contrast to early research in rigid-body collisions suggesting that human physical predictions do not obey ground-truth physics, instead relying on heuristics (Gilden & Proffitt, 1994; Runeson et al., 2000). Overall, our results agree with Bates et al.’s (2015) findings: i.e., ISE predictions entailing the noisy Newton framework outperformed ground-truth and data-driven models in each experiment, further reinforcing the critical role of perceptual noise and physical dynamics in intuitive physics reasoning.

Our results also indicate that people naturally attend to latent attributes when reasoning about familiar and unfamiliar substance states following observation of realistic demonstration animations. Although mental simulation has been demonstrated as a default strategy in other mechanical reasoning tasks (Clement, 1994; Hegarty, 2004), the participants in Schwartz and Blacks (1999) experiments failed to spontaneously represent and simulate physical properties relevant to the judgment task. By designing tasks with regard to the three features outlined in the introduction, our participants were able to mentally simulate dynamic events and did not appear to rely on explicit or heuristic-based reasoning. While the present study indicates a set of simulation-inducing task characteristics, further research should aim to determine specific experimental factors that trigger simulation strategies. Specifically, can the conditions employed in the present tasks extend to classical rigid-body and fluid dynamics problems to resolve the discrepancy between peoples explicit predictions and tacit judgments, and if so, what additional task characteristics serve to facilitate mental simulation?

5.7.1 Are Precise Numerical Simulation Methods Needed to Explain Human Judgments?

Taken together, our results demonstrate that human predictions and judgments about substance dynamics can be accounted for by a unified simulation method with uncertainty implemented into underlying physical variables. However, classical research in artificial intelligence has traditionally dismissed robust mental simulation as a strategy for physical reasoning due to its inherent complexity, often proposing simplified qualitative models in-
stead (De Kleer & Brown, 1984). While the computational substance simulations employed in the present study require extensive numerical evaluation to make predictions about future substance states, humans appear to do so with precision and accuracy in comparatively small amounts of time. It is unlikely that the human brain numerically evaluates partial differential equations to discern whether physical quantities (e.g., mass and momentum) are conserved, nor is it likely that the brain stores the locations of vast numbers of particles to form physical predictions and judgments. Instead, our results provide evidence that humans approximate the dynamics of substances in a manner consistent with ground-truth physics but succumb to biases invoked by perceptual noise when inferring future environmental states. It remains unclear, however, whether the dynamics of rigid objects, liquids, and granular materials can be approximated by a simplified physical model as opposed to the highly complex Navier-Stokes equations solved by the MPM.

To answer this question, we constructed a second simulation-based model which approximates substances as a set of rigid balls which interact with one another according to the principle of conservation of momentum: i.e., the same principles governing the dynamics of the rigid balls in Experiment 3. In the ball approximation model (BAM), latent attributes of liquid and sand were approximated by damping the angular acceleration of each ball with magnitude proportional to its angular velocity, weighted by a stiffness parameter. Similar to noisy viscosity, friction angle, and restitution, noisy stiffness was generated by offsetting a mean value parameter with Gaussian noise on a logarithmic scale. Although the BAM predictions in Experiment 1 were less correlated with human performance than the ISE—\( r(14) = 0.97, \text{RMSD} = 0.09 \)—the approximation model outperformed competing data-driven models.

5.7.2 Implementing Mental Simulation Uncertainty Into the MPM

While human results are generally consistent with the physics-based simulation models coupled with noisy input variables, there remain discrepancies between model predictions and human judgments. Although our ISE and the BAM accounted for perceptual uncertainty in
each situation, the simulations themselves closely approximated normative physical principles. Adding “stochastic noise” to physical dynamics (i.e., mental simulation uncertainty), however, has been shown to increase model performance when predicting human responses in simple physical situations (Smith & Vul, 2013). While mental simulation uncertainty can easily be built into rigid-body collisions, employing this strategy in the present physical simulations would preclude stable numerical evaluation. Since adding noise to the position and velocity of particles in the ISE causes the numerical solutions to diverge, one potential alternative is adding noise to variables upon which the position and velocity depend—namely, the physical attributes specified in each task. However, while it makes sense to suppose that the spatially represented positions or motions of discretized substance particles may change with time, it is unclear whether human estimates of the underlying attributes also fluctuate. Moreover, adding drastic attribute changes to constituent material points throughout time would likely perturb the stability of numerical simulation methods, so an appropriate perturbation method would need to be developed. We hope that future work utilizing substance simulation engines in intuitive physics will pursue this direction of research.

### 5.7.3 Concluding Remarks

The results reported herein demonstrate that human expectations about the dynamics of non-solid substances are consistent with ground-truth physics, given that their representations are prone to observational and physical uncertainty. Moreover, people can predict the future status of situations involving granular materials (i.e., sand), which is less common in daily life than viscous liquids or collections of rigid objects. Results from the BAM also demonstrate the viability of coarse spatial approximation in substance-related situations, suggesting that human observers could be doing something similar when reasoning about the dynamics of liquids and other non-rigid materials.

Recent deep learning architectures in intuitive physics have demonstrated that pattern-recognition networks can learn interaction rules between rigid objects from observation of 2D situations (Battaglia et al., 2016; Chang et al., 2016; Grzeszczuk et al., 1998). A natural
extension of the BAM is to discern whether a deep learning (interaction) network can be trained on corresponding rigid-body approximations of 2D substance dynamics events and tested on new situations. It would be exciting to explore whether interactions between rigid substance elements can be learned and used to approximate substance dynamics in novel situations. It is our hope that future work will proceed in this direction.
6.1 Abstract

Research on analogical problem solving has shown that people often fail to spontaneously notice the relevance of a semantically remote source analog when solving a target problem, although they are able to form mappings and derive inferences when given a hint to recall the source. Relatively little work has investigated possible individual differences that predict spontaneous transfer, or how such differences may interact with interventions that facilitate transfer. In this study, fluid intelligence was measured for participants in an analogical problem-solving task, using an abridged version of the Raven’s Progressive Matrices (RPM) test. In two experiments, we systematically compared the effect of augmenting verbal descriptions of the source with animations or static diagrams. Solution rates to Duncker’s radiation problem were measured across varying source presentation conditions, and participants’ understanding of the relevant source material was assessed. The pattern of transfer was best fit by a moderated mediation model: the positive impact of fluid intelligence on spontaneous transfer was mediated by its influence on source comprehension; however, this path was in turn modulated by provision of a supplemental animation via its influence on comprehension of the source. Animated source depictions were most beneficial in facilitating spontaneous transfer for those participants with low scores on the fluid intelligence measure.
6.2 Introduction

Analogical inference—the application of knowledge about a familiar source system to a novel but structurally similar target system—is critical in scientific discovery (Dunbar & Klahr, 2012) and many other types of creative human activity (Gentner, 2010; Holyoak, 2012; Holyoak & Thagard, 1995). The human capacity for abstract thinking, which is exemplified by analogical reasoning, exceeds that of any other species and plays a significant role in formulating ideas that transcend immediate perception (Penn, Holyoak, & Povinelli, 2008).

It is generally recognized that analogical reasoning involves several sub-processes, most notably retrieval of a related source analog, mapping, inference, and subsequent generalization (e.g., Holyoak, Novick, & Melz, 1994). A number of computational models of these sub-processes have been proposed, including the Structure Mapping Engine (SME; Falkenhainer, Forbus, & Gentner, 1989), the Incremental Analogy Machine (IAM; Keane & Brayshaw, 1988; Keane, Ledgeway, & Duff, 1994), the Structured Tensor Analogue Reasoning model (STAR; Halford, Wilson, & Phillips, 1998), and Learning and Inference with Schemas and Analogies (LISA; Hummel & Holyoak, 1997, 2003). A basic empirical finding is that when a source and target are drawn from different knowledge domains and encountered in different contexts, a potentially useful source analog often remains unnoticed. The gap between noticing and actual use of a source analog has been explored most extensively in experiments using Duncker’s (1945) radiation problem as the target analog. In this problem, a doctor must find a way to use a radiation ray of varying intensity to destroy an inoperable stomach tumor in a patient. The essence of the problem is that high-intensity rays will destroy healthy tissue when they pass through it on their way to the tumor. Although low-intensity rays do not harm healthy tissue, they are also ineffective in damaging or destroying the tumor. The convergence solution is to apply multiple low-intensity rays to the tumor simultaneously from multiple locations surrounding the target.
6.2.1 Spontaneous Analogical Transfer

Gick and Holyoak (1980, 1983) found that in the absence of a related source analog, only about 10% of the participants were able to generate the convergence solution to the radiation problem. When a verbal story highly dissimilar to the radiation problem (a story about a general using converging troops to capture a fortress) was presented to participants prior to the target problem, the rate of spontaneously generating convergence solutions increased to about 30%. After receiving an explicit hint to recall the source analog, an approximately 50% additional participants gave the convergence solution, for a total solution rate of roughly 80%. Thus, people often failed to spontaneously notice the relevance of the source in solving the target problem, though they could successfully form mappings and derive inferences when prompted to do so. The difficulty of spontaneously noticing the relevance of distant analogs remains even when more naturalistic materials are employed (Trench & Minervino, 2015).

Subsequent research has established that spontaneous transfer can be facilitated in a number of ways, including choosing a source analog that is relatively similar to the target (Keane, 1987), or one that permits a clear, isomorphic mapping to the target problem (Holyoak & Koh, 1987). In addition, close comparison of multiple source analogs appears to aid in abstracting a more general schema for a class of problems, which in turn fosters later spontaneous transfer (Catrambone & Holyoak, 1989; Gick & Holyoak, 1983; Loewenstein, Thompson, & Gentner, 2003). Direct instructions to search for remote analogs (without specifying any specific domain) can also be effective (Trench, Olguín, & Minervino, 2016). In general, manipulations that encourage attention to shared relations tend to foster transfer of learning (for a review, see Goldwater & Schalk, 2016).

Less work has been done to address the possibility that people systematically differ in their ability to spontaneously notice and use remote analogies. Evidence shows that measures of fluid intelligence, notably Raven’s Progressive Matrices (Raven, 1938), predict performance on standardized analogy tests (Snow, Kyllonen, & Marshalek, 1984) as well as analogical mapping in experimental tasks (Vendetti, Wu, & Holyoak, 2014). Fluid in-
telligence is closely linked to working-memory capacity (Ackerman, Beier, & Boyle, 2005; Engle, 2002), which has been linked to analogical reasoning in some computational models (e.g., Halford et al., 1998; Hummel & Holyoak, 2003). However, most tests of analogical reasoning present the source and target together (typically in A:B :: C:D format), so that the need for spontaneous retrieval of the source is eliminated. Relatively few studies have examined individual differences in more complex analogical problem solving (e.g., Antonietti & Gioletta, 1995). There is some evidence that verbal ability predicts spontaneous transfer rates when the source analog is relatively unfamiliar (Corkill & Fager, 1995), and that a measure of math expertise (math SAT score) predicts spontaneous transfer to a math generalization problem after exposure to multiple target analogs (Novick & Holyoak, 1991).

This study is (to our knowledge) the first to investigate whether fluid intelligence scores predict spontaneous analogical transfer. As noted, computational models of analogical reasoning that operate within working-memory constraints imply that the mapping process will be sensitive to working-memory capacity. However, mapping necessarily follows the process of accessing a source analog in memory (since by definition mapping must consider the source and target analogs together). It is therefore less clear on theoretical grounds whether working-memory capacity, or the closely related concept of fluid intelligence, will have an impact on retrieval of the source analog, and hence on spontaneous transfer.

We further aimed to determine whether the impact of fluid intelligence on spontaneous transfer is mediated by variations in comprehension of the source analog. In addition, we sought to determine whether the influence of fluid intelligence is modulated by an important factor that may facilitate analogical transfer: providing animations to support comprehension of the source analog.

6.2.2 Visuospatial Displays and Analogical Transfer

Research on diagrammatic reasoning has shown that visuospatial representations of solution strategies for mechanical problems can enhance people’s ability to infer the principles of operation for physical systems (Hegarty & Stull, 2012), suggesting the importance of display
format in acquisition of abstract knowledge. A few studies have shown that static, visual diagrams can be used as source analogs for verbal target problems (Gick, 1985; Gick & Holyoak, 1983). While uninterpreted diagrams generally result in low rates of spontaneous transfer, they can serve as effective analogs following a hint to recall and apply them to a novel target problem (Gick & Holyoak, 1980).

In contrast to the weak effects of static diagrams, there is some evidence that animated displays can facilitate spontaneous transfer (Beveridge & Parkins, 1987; Pedone et al., 2001). The radiation problem is temporally dynamic, in that the key concepts involve the summation of forces over space and time. The use of physical motion in an animated display may help the learner to focus attention on dynamic relationships (Tversky & Morrison, 2002), which may in turn provide additional retrieval pathways when the target problem is encountered. Day and Goldstone (2011) found that presenting a force-based physical system can prime dynamic mental models, which in turn facilitates spontaneous transfer when solving social problems based on superficially dissimilar dynamic systems.

The potential effectiveness of animation in promoting transfer is consistent with work in fields such as math education, where it has been advocated that students should be encouraged to ground formalisms in concrete situation models: i.e., approximate perceptual and dynamic representations of how events described by a problem text occur in the real world (Koedinger, Anderson, Hadley, & Mark, 1997; Nathan, 1998; Nathan, Kintsch, & Young, 1992; Reed, 2006). Simulations based on situation models have proved effective in promoting better comprehension of relatively simple texts (e.g., Glenberg, Gutierrez, Levin, Japuntich, & Kaschak, 2004). An important finding is that far transfer is often best promoted by using relatively idealized representations, such as simplified rather than detailed perceptual displays (for a review, see Fyfe, McNeil, Son, & Goldstone, 2014).

This study aimed to systematically compare the effectiveness of animations and static diagrams (combined with verbal descriptions) in facilitating spontaneous analogical transfer to the radiation problem. In a previous investigation of the impact of animation on analogical transfer, Pedone, Hummel, and Holyoak (2001) presented animations and diagrams without any verbal cover story and did not measure participants’ understanding of the source
analogs. Evidence shows that the combination of animations and spoken narration is especially effective in increasing understanding of a mechanical system (Mayer, 2009; Mayer & Anderson, 1991). Thus, animations may provide deeper insight into the causal structure of a dynamic system than does a verbal description alone. To assess this possibility, our study included measures of participants’ understanding of the source analog. We hypothesized that animations would improve initial understanding of the source and facilitate subsequent spontaneous analogical transfer. The animations we tested were relatively simple, aiming to focus attention on key relations while grounding verbal problem descriptions in a dynamic perceptual representation. By measuring individual differences, we also sought to determine whether the impact of animation on spontaneous transfer differs for people at varying levels of fluid intelligence.

6.3 Experiment 1: Animated Source Materials Facilitate Transfer

6.3.1 Method

Participants. A total of 126 (92 female; N = 42 per condition) participants were recruited from the Department of Psychology subject pool at the University of California, Los Angeles (UCLA). They were compensated with course credit for their participation and were naïve to the purpose of the experiment. Three participants with less-than-chance performance on the fluid intelligence measure (Raven’s score < 2) were removed from the analysis.

Source Analog Stimuli. Animated stimuli were generated using Psychophysics Toolbox Version 3 (Brainard, 1997; Kleiner et al., 2007) in MATLAB, and presented on a CRT monitor at a viewing distance of 70 cm. In each animation, either one or eight cannons fired small or large cannonballs radially inward toward an enemy octagon surrounded by a friendly barrier. The octagon was colored red and had an initial angular width of 0.89. In the single-cannon scenarios, the octagon’s angular width was reduced by 0° and 0.007° after each small- and large-cannonball volley, respectively. In the multiple-cannon scenarios, its width was reduced by 0.040° and 0.056° after each volley. The small and large cannonballs
Table 6.1: **Cannon Scenario Attributes.** Number of cannons, size of cannonballs, and level of damage inflicted to key elements in each scenario.

<table>
<thead>
<tr>
<th>Scenario Number</th>
<th>Number of Cannons</th>
<th>Cannonball Size</th>
<th>Barrier Damage</th>
<th>Octagon Damage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Small</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Large</td>
<td>Major</td>
<td>Minor</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>Large</td>
<td>Major</td>
<td>Major</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>Small</td>
<td>None</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

were colored white and had angular widths of 0.10° and 0.20°, respectively. Each cannon was also colored white and had an angular length and width of 1.13° and 0.57°, respectively. The barrier was colored green and had a radial width that subtended 0.25° of visual angle.

Upon impact of each large cannonball, the outer radius of the barrier was reduced by a Gaussian function with maximum magnitude at the point of impact and decreasing magnitude with azimuthal distance. The outer radius was reduced until the barrier was breached, which occurred after 10 large-cannonball volleys. The minimal radial distance to the barrier and each cannon subtended 1.19° and 4.15° of visual angle, respectively. Following each scenario, a message box appeared, indicating whether the cannon(s) succeeded or failed to meet their objective (i.e., whether the enemy octagon was defeated). Elements in each animation were displayed over a black background. Figure 6.1 illustrates representative diagrams for each scenario, which were chosen from intermediate frames from their respective animations. The number of cannons and the size of the cannonballs in each scenario, along with the level of damage inflicted to the key elements in each display, are indicated in Table 6.1. The depicted frames were chosen such that the amount of damage inflicted to the barrier and octagon were visually apparent. Videos of the animated scenarios, along with their accompanying spoken monologues, are included in online Supplemental Materials (http://cvl.psych.ucla.edu/moviedemo.html).
Figure 6.1: **Illustration of Scenario Diagrams in the Diagram + Verbal Condition.** Participants were presented with the scenarios in sequential order, accompanied by an auditory-verbal explanation of key concepts. (a) In Scenario 1, a single cannon fires small cannonballs and inflicts no damage to either the barrier or octagon. (b) In Scenario 2, a single cannon fires large cannonballs and inflicts minor damage to the octagon and major damage to the barrier. (c) In Scenario 3, multiple cannons fire large cannonballs simultaneously, this time inflicting major damage to both the octagon and barrier. (d) In Scenario 4, multiple cannons fire small cannonballs simultaneously. The converging cannonballs inflict no damage to the barrier and moderate damage to the octagon.

Both animations and diagrams were accompanied by spoken monologues recorded on a Blue Yeti USB microphone with a 48 kHz (16 bit) recording sample rate. In each spoken monologue, the number of cannons and the size of the cannonballs were specified in addition to the amount of damage inflicted to the friendly barrier and enemy octagon by the cannon(s). There were four possible levels of damage (none, minor, moderate, and major), which varied across scenarios. The same source analog stimuli were employed in both Experiments 1 and 2 of this study.
Procedure and Design. Participants were randomly assigned into one of three conditions (Verbal, Diagram + Verbal, and Animation + Verbal) reflecting the presentation method for the source analog. In all conditions, an initial instruction page provided pictorial depictions of the general elements in the source analog, which was comprised of four scenarios presented in sequential order. Participants in each condition received an auditory-verbal (spoken monologue) description of each scenario. Those in the Verbal condition were presented with the spoken monologue alone, whereas participants in the Diagram + Verbal and Animation + Verbal conditions received a supplemental diagram or animation, respectively.

The goal of the cannon(s) in each of the four scenarios is to defeat the enemy octagon without inflicting critical damage to any local region of the barrier. The successful scenario is analogous to the convergence solution to Duncker’s radiation problem, in which multiple radiation sources are fired at low intensity from multiple locations surrounding the patient’s stomach, successfully destroying the tumor without inflicting critical damage to any local region of the surrounding healthy tissue. The relational structure of the cannonball scenario is thus isomorphic to the key relations in the potential solution to the radiation problem. Questions related to each scenario were presented sequentially following the presentation of all four scenarios. Participants answered two multiple-choice questions per scenario, which assessed their understanding of the level of damage inflicted to the barrier and octagon in each system (e.g., “What level of damage did the cannonballs cause to the friendly barrier in Scenario 1?”). Damage level was directly stated in the spoken monologue for each scenario. After completing the multiple-choice questions, participants were asked to explain why the cannon(s) failed or succeeded in meeting the objective in each scenario (one free-response question per scenario). The question materials are provided in the Appendix. Both multiple-choice and free-response questions were administered using Qualtrics, an online survey environment intended for research and experimental purposes.

Next, each subject completed an abridged, 12-item version of the Raven’s Progressive Matrices test (RPM; Arthur Jr., Tubré, Paul, & Sanchez-Ku, 1999). In this task, participants viewed a series of $3 \times 3$ matrices with textured shapes displayed in each cell. The shapes exhibited systematic patterns across the rows and columns of each matrix, and participants
were instructed to identify the missing element that completed the pattern without violating the underlying relational rules. The test provides a nonverbal measure of fluid intelligence, allowing us to assess potential individual differences in transfer performance. The abridged RPM test also served as a filler task to create a total delay (approximately 15 minutes) between presentation of the source and target analogs. Participants required approximately 11 minutes to complete the RPM test.

Finally, transfer rates to the radiation problem (Duncker, 1945; included in Appendix D) were measured across varying conditions of source training. Participants were asked to solve Duncker’s radiation problem in a two-pass fashion (cf. Gick & Holyoak, 1980). On the first pass, participants received no indication that the previously presented scenarios were related to the target problem. Spontaneous transfer was defined by successful generation of the convergence solution on the first pass without any hint. After participants submitted their answers, the radiation problem was presented again, but this time with an explicit hint prompting participants to recall the cannon scenarios and any solution to the radiation problem they might suggest. Hence, hinted transfer was defined by successful generation of the convergence solution on the second pass, after the explicit hint.

6.3.2 Results

**Analogical Transfer Rates.** A set of criteria (adapted from previous research; Gick & Holyoak, 1980, 1983) were used to determine whether participants successfully solved the radiation problem either before or after the hint. Solutions were scored according to whether participants conveyed at least two of the three critical ideas underlying the convergence principle (i.e., multiple radiation sources are needed, radiation sources should fire low-intensity rays, and radiation sources should be positioned in different locations surrounding the patient’s stomach tumor). Participants were scored as having either (i) solved the radiation problem spontaneously (i.e., without the hint), (ii) solved the radiation problem with the hint, or (iii) failed to solve the radiation problem. Two undergraduate research assistants naïve to the experimental hypothesis and test conditions independently scored each partici-
Figure 6.2: **Spontaneous and Total Transfer Rates.** Transfer rates (percentage of participants who successfully solved the radiation problem) for the Verbal, Diagram + Verbal, and Animation + Verbal conditions are displayed (Experiment 1).

Participants’ responses, with an agreement rate of 95% (Cohen’s $\kappa = .87$). A third researcher broke the tie if the first two researchers disagreed with one another.

Spontaneous transfer rate corresponds to the percentage of participants who produced the convergence solution to the radiation problem before a hint was given to recall the source analog. This measure assesses participants’ ability to spontaneously retrieve the source analog and apply their knowledge to a novel target problem following a time delay. Total transfer rate corresponds to the percentage of participants able to solve the radiation problem either before or after a hint was given. Figure 6.2 depicts percentage of spontaneous transfer before a hint, and total transfer percentage after a hint, for each condition (42 participants per condition).

Spontaneous transfer rates were 55%, 50%, and 83% for the Verbal, Diagram + Verbal, and Animation + Verbal conditions, respectively. The spontaneous transfer rate in the Animation + Verbal condition was reliably greater than that obtained in the Diagram + Verbal condition—$\chi^2(1) = 10.50, \phi = .35, p < .01$—as well as the Verbal condition—$\chi^2(1) = 8.02, \phi = .31, p < .01$—indicating that animated source instruction facilitated spontaneous analogical inference. Spontaneous transfer rate did not differ significantly between the Verbal...
and Diagram + Verbal conditions—χ²(1) < 1, φ = .05—suggesting that the addition of a static pictorial display was not effective in priming the temporally dynamic convergence principle.

Total transfer rates after the hint were 83%, 69%, and 88% for the Verbal, Diagram + Verbal, and Animation + Verbal conditions, respectively. This level of overall performance is consistent with previous findings showing that roughly 80% of college students are able to solve the radiation problem following a hint to think back to a relevant source analog (Gick & Holyoak, 1980, 1983). The total transfer rate in the Animation + Verbal condition exceeded that obtained in the Diagram + Verbal condition, χ²(1) = 4.53, φ = .23, p = .03, but did not differ from that obtained in the Verbal condition, χ²(1) < 1, φ = .07. The total transfer rate in the Diagram + Verbal condition was not statistically different from that in the Verbal condition, χ²(1) = 2.36, φ = .17, p = .12.

**Comprehension of Source Analog.** To assess whether the advantage of the animated source depiction in supporting spontaneous transfer was linked to deeper understanding of the source analog, we evaluated participants’ responses to multiple-choice (MC) and free-response (FR) source-understanding questions. The eight MC questions aimed to measure participants’ understanding of how the small and large cannonballs interact with the various elements in the system, and how their forces sum together across scenarios. Each MC question was scored as either correct or incorrect, according to whether the participant selected the correct amount of damage inflicted to the specified element by the cannon(s). The maximum score for MC questions was thus 8 points.

The FR questions assessed participants’ understanding of why the cannon(s) failed or succeeded in each scenario. For FR responses, three key principles were chosen for each scenario (see Appendix D). Participants received one point for each correctly conveyed principle and had one point deducted for each incorrect idea they stated. For each question, participants receiving 2 or more points were given a score of 2, participants with one point were given a score of 1, and participants with zero points or less were given a score of 0. The maximum score on the four FR questions was thus 8 points (the same as for MC questions). The FR responses in each scenario were scored by two researchers. If two of the three scorers
agreed, their score was used. A third scorer was employed for those responses where the first two disagreed. If the three scorers disagreed, the response was jointly discussed until two researchers agreed on a score. The agreement rate for the first two scorers was 80% (Cohen’s $\kappa = .62$). The agreement rate for three scorers (i.e., cases where two of the three scorers agreed) was 95%.

FR and MC scores were correlated across participants ($r = .47$, $p < .001$). For FR scores, participants in the Animation + Verbal condition scored highest, $t(82) = 2.27$, $d = .50$, $p = .03$, Animation versus Diagram, followed by the Diagram + Verbal condition, $t(82) = 2.23$, $d = .49$, $p = .03$, Diagram versus Verbal, and the Verbal condition scored the lowest. For MC scores, participants in the Animation + Verbal condition again scored higher than those in either the Diagram + Verbal, $t(82) = 2.76$, $d = .60$, $p < .01$, or Verbal conditions, $t(82) = 3.07$, $d = .67$, $p < .01$, whereas there was no significant score difference between Diagram + Verbal and Verbal participants, $t(82) < 1$, $d = .11$. The consistent superiority of the animation condition in both FR and MC responses indicates that presence of a supplemental animation promoted deeper understanding of the source analog, relative to other presentation methods (see Figure 6.3).

**Individual Differences in Fluid Intelligence in Relation to Spontaneous Analogical Transfer.** To determine whether fluid intelligence influences spontaneous analogical transfer and its interaction with presentation format of the source analog, we performed a median split on participants according to their Raven’s score, classifying them as either low (Raven’s score $< 8$) or high (Raven’s score $> 8$) on the RPM test. Forty-five participants were classified as low RPM ($N = 15, 12, and 18$ for the Verbal, Diagram + Verbal, and Animation + Verbal conditions, respectively), and 57 participants were classified as high RPM ($N = 19$ for each condition). The other 24 participants, who had median Raven’s scores of 8, were dropped from the individual differences analysis as they could not be reasonably classified as either high or low. Data from those participants with median Raven’s scores were, however, used in a subsequent moderated mediation analysis, which aggregated data from Experiments 1 and 2 (see Moderated Mediation Analysis section).
Figure 6.3: Free-Response (FR) and Multiple-Choice (MC) Source-Understanding Scores. Average scores are displayed for participants in the Verbal, Diagram + Verbal, and Animation + Verbal conditions (Experiment 1). Error bars indicate ± 1 standard error of the mean (SEM).

We first compared free-response (FR) and multiple-choice (MC) source-understanding scores between the low- and high-RPM groups. Both FR and MC scores in the high-RPM group exceeded those in the low-RPM group—\( t(100) = 3.66, d = .73, p < .001 \) for FR scores; \( t(100) = 3.83, d = .76, p < .001 \) for MC scores—indicating superior source understanding for participants scoring relatively high on our fluid intelligence measure. We then compared FR and MC scores between conditions for the low- and high-RPM groups (see Figure 6.4). For low-RPM participants, both FR and MC scores in the Animation + Verbal condition exceeded those in the Diagram + Verbal condition—\( t(28) = 3.79, d = 1.41, p < .001 \) for FR scores; \( t(28) = 2.09, d = .78, p < .05 \) for MC scores—and Verbal condition—\( t(31) = 3.92, d = 1.37, p < .001 \) for FR scores; \( t(31) = 4.37, d = 1.53, p < .001 \) for MC scores—indicating enhanced source understanding for low-RPM participants when the source analog included an animation. For high-RPM participants, MC scores did not differ reliably across the three conditions. However, FR scores in the Animation + Verbal condition exceeded those in the Verbal condition, \( t(36) = 2.25, d = .73, p = .03 \), although no difference was observed between FR scores in the Animation + Verbal and Diagram + Verbal conditions, \( t(36) < 1, d = .12 \).
Figure 6.4: **Free-Response (FR) and Multiple-Choice (MC) Source Understanding Scores Across RPM Group.** Average scores are shown for low-RPM (left) and high-RPM (right) participants across experimental conditions (Experiment 1). Error bars indicate ± 1 SEM.

FR scores in the Diagram + Verbal condition also exceeded those obtained in the Verbal condition, $t(36) = 2.03, d = .66, p < .05$, indicating that static diagrams were sufficient to facilitate source understanding (measured by the FR metric) for high-RPM participants.

Spontaneous transfer rates were then compared across RPM groups, revealing superior performance for those participants scoring high on the fluid intelligence measure, $\chi^2(1) = 4.47, \phi = .21, p = .03$. (Because total transfer rates after a hint approached the effective ceiling level, analyses of individual differences were not performed for that measure.) Next, spontaneous transfer rates were compared across conditions for the low- and high-RPM groups (see Figure 6.5). For low-RPM participants, the spontaneous transfer rate for the radiation problem in the Animation + Verbal condition exceeded that obtained in the Diagram + Verbal, $\chi^2(1) = 7.75, \phi = .51, p < .01$, and Verbal, $\chi^2(1) = 6.64, \phi = .45, p = .03$, conditions. No difference was observed in transfer performance between the Diagram + Verbal and Verbal conditions, $\chi^2(1) < 1, \phi = .07$. For high-RPM participants, the spontaneous transfer rate in the Animation + Verbal condition was marginally greater than that in the
Figure 6.5: **Spontaneous Transfer Rate Across RPM Group.** The percentage of convergence solutions generated before hint for low-RPM (left) and high-RPM (right) participants are displayed across experimental conditions (Experiment 1)

Diagram + Verbal condition, $\chi^2(1) = 3.64$, $\phi = .31$, $p = .06$, but did not differ from that in the Verbal condition, $\chi^2(1) = 1.58$, $\phi = .20$, $p = .21$. The spontaneous transfer rate in the Diagram + Verbal condition also did not differ from that in the Verbal condition, $\chi^2(1) < 1$, $\phi = .11$.

In summary, animated source depictions were most beneficial in fostering comprehension of the source analog and spontaneous analogical transfer for those participants with low scores on the fluid intelligence measure. After describing Experiment 2, we will report mediation analyses that further gauge the relationships among animation, source understanding, and spontaneous transfer in analogical discovery.

### 6.4 Experiment 2: Source Understanding Questions Inhibit Transfer

The results of Experiment 1 indicate that the advantage of an animated source analog in supporting spontaneous transfer was linked to greater understanding of the source scenarios,
measured by our MC and FR questions. One issue that remains open is what, if any, impact the answering of questions about the source may have itself had on spontaneous transfer. The questions provided a measure of source understanding that proved to be predictive of spontaneous transfer. However, to establish a baseline for comparing transfer rates in our studies to those in related studies that did not administer such questions, Experiment 2 explicitly varied whether or not participants answered questions about the source analog.

It is unclear whether answering such questions itself facilitates or impedes transfer, compared to previous studies in which participants’ understanding of source analogs was not measured prior to presentation of the radiation problem (e.g., Gick & Holyoak, 1980, 1983; Pedone et al., 2001). On the one hand, responding to questions about the cannon scenarios could have helped participants attend to the abstract principle as instantiated in the source analog, thereby strengthening potential retrieval cues for subsequent recall and transfer to the radiation problem. On the other hand, the questions probed domain-specific details about the source analog—unlike questions that elicit comparisons of multiple analogs to one another (cf. Gick & Holyoak, 1980). Although performance on the scenario-understanding questions predicted abstract understanding of the underlying principle, it is possible that answering these questions may have strengthened memory for scenario-specific details at a featural level that would not be useful for solving the radiation problem. Thus, the aim of Experiment 2 was to determine what effect, if any, administering source-understanding questions has on analogical transfer to the radiation problem. In addition, we wanted to increase the overall power of the basic design to obtain enough data to perform a mediation analysis.

6.4.1 Method

Participants. A total of 240 participants (159 female; $N = 60$ per condition) were recruited from the Department of Psychology subject pool at UCLA. They were compensated with course credit for their participation and were naïve to the purpose of the experiment. One participant with less-than-chance performance on the fluid intelligence measure (Raven’s score $< 2$) was removed from the analysis.
**Procedure and Design.** The Diagram + Verbal condition was dropped in Experiment 2, leaving two levels for the between-subjects factor of source presentation format (Verbal and Animation + Verbal). Participants were randomly assigned into two conditions of a second between-subjects factor (Questions and No Questions), reflecting whether or not source-understanding questions were asked following presentation of the source scenarios. We recognized that eliminating the questions would decrease the time delay between presentation of the source scenarios and the radiation problem. Since participants in Experiment 1 spent approximately 218 seconds answering source-understanding questions, a delay of 18 seconds was introduced following each RPM item for participants in the No Questions condition (yielding a total delay of approximately 15 minutes). The time required for participants to complete the experiment did not differ between the Questions and No-Questions groups, $t(238) = .48, d = .06, p = .64$, nor did participants’ Raven’s scores, $t(238) = 1.27, d = .16, p = .21$. In all other respects the procedure and design were identical to those of Experiment 1.

### 6.4.2 Results

**Analogical Transfer Rates.** The criteria for scoring participants’ responses to the radiation problem were the same as in Experiment 1. The agreement rate between the two researchers was 88% (Cohen’s $\kappa = .77$), and a third researcher again broke the tie if the first two researchers scores disagreed with one another. Because the interaction between source presentation format and presence of questions was not significant in a binary logistic regression model predicting spontaneous and total transfer, we proceeded to interpret the main effects of source presentation format and asking source-understanding questions.

Replicating the major finding of Experiment 1, both spontaneous and total transfer rates in the Animation + Verbal condition exceeded those in the Verbal condition, 83% versus 58%, $\chi^2(1) = 18.15, \phi = .28, p < .001$, for spontaneous transfer; 90% versus 73%, $\chi^2(1) = 12.06, \phi = .22, p < .001$, for total transfer. Spontaneous transfer rates were 63% in the Questions condition and 79% in the No-Questions condition, indicating that the presence
of source-understanding questions appeared to impede spontaneous transfer performance, \( \chi^2(1) = 8.07, \phi = .18, p < .01 \). Total transfer rates did not differ significantly between the No-Questions and Questions conditions, 85\% versus 78\%, \( \chi^2(1) = 2.22, \phi = .10, p = .14 \). These findings suggest that answering questions about the source scenarios may strengthen memory for source-specific information that is not helpful for spontaneously transferring abstract knowledge between the source analog and the radiation problem. Such source-specific information would increase the semantic distance between the source and target, increasing the difficulty of subsequent transfer. Consequently, the net effect of asking questions was to impair spontaneous retrieval of the source.

**Moderated Mediation Analysis.** To assess the impact of individual differences in fluid intelligence on spontaneous transfer, directly and via an influence on source comprehension—and how animation modulates these relationships—we conducted mediation and moderation analyses on combined data from Experiment 1 and the Questions condition of Experiment 2. Hence, the analysis was based on observations from a total of 204 participants. These analyses treated fluid intelligence, as measured by RPM score, as a continuous variable, thereby avoiding the pitfalls associated with creating a binary dichotomous variable (i.e., classifying participants as either low or high RPM). A single source-understanding score was calculated for each participant by summing together FR and MC scores, and a dummy coded variable was created to indicate presence of animation (1 if animations were present, 0 if absent). Both RPM and summed source-understanding scores were standardized prior to the mediation analysis.

As shown in Figure 6.6a, the relationship between fluid intelligence and spontaneous analogical transfer was mediated by source understanding. Moreover, the indirect effect of fluid intelligence on spontaneous transfer through source understanding was moderated by provision of a supplemental animation with the source analog. We compared this moderated mediation model to a baseline mediation model including an indirect effect of presence of animation on spontaneous transfer through source understanding, but omitting the influence of individual differences in fluid intelligence (see Figure 6.6b). Log likelihoods were calculated.
Figure 6.6: **Moderated Mediation Analysis (Experiments 1 and 2).** Standardized regression coefficients for a) the indirect effect of Raven’s score on spontaneous transfer through source understanding moderated by presence of animation; and b) the indirect effect of presence of animation on spontaneous transfer through source understanding. The log likelihood \((LL)\) value indicates the goodness of fit for each model; values closer to zero correspond to higher likelihoods using predicted source-understanding scores, and a likelihood ratio test was conducted to compare the goodness of fit for each model. Results revealed a superior goodness of fit for the moderated mediation model that included the individual difference variable, \(\chi^2(2) = 14.84\), Cramer’s \(V(1) = .35\), \(p < .001\).

We tested the significance of links in the baseline and moderated mediation models using bootstrapping procedures. Standardized indirect effects were computed for each of 50,000
bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. If the bootstrapped confidence interval does not include zero, the corresponding effect is significant at $p = .05$. For the baseline mediation model, the bootstrapped standardized indirect effect was statistically significant ($ab = .80, CI_{95} = [.45, 1.28]$), as was the relationship between source understanding and spontaneous transfer ($b = 1.00, CI_{95} = [.59, 1.40]$). The relationship between presence of animation and source understanding was also statistically significant ($a = .81, CI_{95} = [.55, 1.06]$), whereas the direct effect of presence of animation on spontaneous transfer was not ($c' = .54, CI_{95} = [-.15, 1.23]$).

For the moderated mediation model, the bootstrapped standardized indirect effects in the presence and absence of animation were statistically significant ($ab = .20, CI_{95} = [.06, .39]$ and $ab = .61, CI_{95} = [.31, 1.00]$, respectively). The relationship between source understanding and spontaneous transfer rate was also significant ($b = 1.01, CI_{95} = [.61, 1.41]$). However, the direct effect of Raven’s score on spontaneous transfer was not reliable ($c' = .29, CI_{95} = [-.06, .64]$). To test whether presence of animation significantly moderated this mediation effect, we calculated the 95% confidence interval of the standardized index of moderated mediation (i.e., the difference between the indirect effects in the presence and absence of animation). We found that the standardized index of moderated mediation was statistically significant ($\Delta ab = .40, CI_{95} = [.12, .79]$), indicating that the indirect effect of fluid intelligence on spontaneous transfer through source understanding was moderated by the presence of animation.

In Experiment 1, we found that supplemental animations were most beneficial for those participants scoring low on the fluid intelligence measure, suggesting a stronger relationship between Raven’s score and source understanding in the absence of animation. Results from the moderated mediation analysis indicate that both presence of animation and Raven’s score had a significantly positive impact on source understanding ($a_{\text{Animation}} = .74, CI_{95} = [.52, .97]$ and $a_{\text{Ravens}} = .60, CI_{95} = [.43, .77]$). However, the impact of Raven’s score on source understanding was attenuated by the presence of animation, as demonstrated by the significant interaction term between the two variables ($a_{\text{Animation}} = -.40, CI_{95} = [-.15, 1.23]$).
Since source understanding predicts spontaneous transfer, we would expect the indirect effect of Raven’s score through source understanding to be greatest in the absence of animation, which is supported by the significant index of moderated mediation. These results provide converging evidence that provision of supplemental animations facilitates spontaneous transfer by increasing comprehension of source materials, and that this benefit is greatest for participants scoring relatively low on the measure of fluid intelligence.

6.5 General Discussion

In two experiments, we found that animated source analogs yielded greater spontaneous transfer than either diagrammatic or purely verbal source analogs. These findings are consistent with those of Beveridge and Parkins (1987) and Pedone, Hummel, and Holyoak (2001). This study went beyond prior work by measuring the impact of both individual differences in fluid intelligence, and varying presentation conditions, on source understanding. We found that providing a supplemental animation of the source analog led to superior source understanding as well as greater spontaneous analogical transfer. In addition, the presence of animation was especially beneficial for participants who scored lower on a measure of fluid intelligence. To the best of our knowledge, this is the first demonstration that fluid intelligence is a predictor of spontaneous analogical transfer (as opposed to prompted analogical mapping; Vendetti et al., 2014).

Animation may aid transfer in at least two ways. First, animated displays serve to ground a verbal problem in a dynamic perceptual experience, creating the benefits associated with multimedia instruction (e.g., Reed, 2006). Second, if the animated display is relatively sparse with respect to visual details, then it may focus attention selectively on the relational aspects of the source analog (Goldwater & Schalk, 2016). In the latter respect, the benefit of animation in promoting an abstract understanding of the source may be similar to the benefit of comparing multiple source analogs, which appears to foster induction of a more abstract schema (Gick & Holyoak, 1983; Novick & Holyoak, 1991). Viewing supplemental animated displays—like comparison of multiple source analogs—may foster richer relational
encodings (Pedone et al., 2001). The dynamic nature of animation may be especially effective in inducing a schema for problems involving the application of forces over time and space. It is possible that an animated source induces internal scan patterns across a mental image of the target problem, which in turn facilitates transfer (analogous to benefits conveyed by certain patterns of eye movements across a diagram of the radiation problem; Grant & Spivey, 2003; Thomas & Lleras, 2007).

Of course, we have only explored the impact of one animated display. Future work should systematically vary the visual richness of such displays to assess whether sparse displays are especially effective in promoting transfer (cf. Fyfe et al., 2014). It is also of interest that the overall rate of spontaneous transfer was substantially higher for the source used in the present study (even in the Verbal-only condition) than in studies using the military source analog (Gick & Holyoak, 1980, 1983). Although there are many differences between the two source analogs, one that may be especially important is that relative to the military story, the “cannonball” scenario used here is less rich in details irrelevant to the convergence solution. Future research could fruitfully examine the general role of degree of specific detail as a factor that influences analogical transfer.

A moderated mediation analysis revealed that fluid intelligence influences spontaneous transfer indirectly, by increasing source comprehension. A possible mechanism is that participants with higher Raven’s scores tend to focus more attention on the underlying relations in the source scenario, essentially adopting a “relational set” (Vendetti et al., 2014). The result is a deeper understanding of the source, based on relational concepts that provide potential retrieval paths once the target problem is encountered. This explanation of how fluid intelligence may facilitate spontaneous analogical transfer is consistent with recent evidence that fluid intelligence as measured by the Raven’s Progressive Matrices test yields benefits that are largely mediated by the use of effective cognitive strategies (Dunlosky & Kane, 2007; Gonthier & Thomassin, 2015).

Animation reduces the impact of fluid intelligence on source comprehension, as it enables those with lower fluid intelligence scores to “catch up” in their source understanding. Our findings suggest that dynamic displays may be especially useful in teaching relational
concepts to students at lower ability levels. Providing these students with more robust representations of a novel source analog may reduce working-memory demands and encourage a relational set, thereby fostering analogical transfer.

Future theoretical work should focus on how dynamic schemas can be learned from animations, perhaps by extending current theories of relation learning (e.g., Lu, Chen, & Holyoak, 2012) to analogical transfer. In addition, across a broad range of topics, the potential value of animations as a teaching tool that can be effectively coupled with analogical examples deserves to be explored further.
CHAPTER 7

Conclusion

Taken together, the preceding chapters constitute progress towards understanding the perception and reasoning components underlying intuitive physics and analogical transfer. In Chapter 2, the current state of the field of intuitive physics was presented in the context of earlier work, outlining new approaches and ongoing research directions. While recent studies have provided evidence that people’s implicit predictions about the physical behavior of objects and substances are consistent with Newtonian principles, it remains unclear how the human brain learns and applies those principles to real-world situations. Although computational approaches utilizing Bayesian statistics and machine learning have shown promise when applied to modern problem domains (e.g., Chapters 3 and 5), successful architectures have mostly relied on generative (Newtonian) models with physical combination rules “written in”. Although deep learning approaches have shown promise in learning physical relations from scratch in 2D worlds (Battaglia et al., 2016; Chang et al., 2016; Grzeszczuk et al., 1998), further work is needed to develop a system which can abstract and transfer knowledge in real-world scenes. I will discuss this issue further in Section 7.1.

There also remains ambiguity in how scenes should be represented when applying learned physical models. As outlined in Section 1.1, situations appear qualitatively different depending on the coordinate axes used to represent the positions of their entities. To date, models under the noisy Newton framework have chosen stationary, Euclidean coordinate axes—or frames of reference—when representing object and substance positions in dynamic scenes. This is a viable choice for toy worlds involving stacked blocks (Battaglia et al., 2013; Hamrick et al., 2016) or poured liquids (Bates et al., 2015; Kubricht et al., 2016) but breaks down for real-world situations where physical entities move in relation to meaningful landmarks in
space and time. Chapter 3 demonstrated that the reasoning system appears to utilize relative motion signals when inferring physical attributes in dynamic (collision) events, which is consistent with earlier findings suggesting that the motion of objects relative to one another is extracted early in the motion perception process (Proffitt et al., 1979). On the face of it, the choice of reference frame seems arbitrary; after all, momentum and energy are conserved as long as positions and motions are represented with respect to non-accelerating reference frames. However, the noisy Newton framework provides the key insight that observable information cannot be directly perceived and must instead be inferred, which relies on prior beliefs about corresponding values. Critically, those prior expectations (e.g., the slow and smooth prior for local motion signals; Weiss & Adelson, 1998) oftentimes need reference points. Objects are expected to move slowly, but in relation to what? The object collision results reported herein suggest that people expect objects to move slowly relative to their local environment: e.g., the surface (background) upon which they move. Designing machines which also encode relative motion information suggests two key benefits. First, relative motion signals are more robust in revealing types of motion than absolute motion signals. For example, a rolling wheel will look roughly the same regardless of its speed or direction of movement when broken down into relative motion components, whereas a point on the wheel’s outer rim will follow different cycloid paths depending on the situation. Second, classifying points in space as meaningful reference points affords a consideration of local context when inferring physical characteristics of observed situations. Rather than needlessly spending energy and computational resources attaining precise estimates of position and velocity in 3D space, Bayesian inference can be leveraged to build an efficient and cost-effective approximation of physical reality. It is my opinion that modeling computational architectures based on characteristics of human perception (e.g., relative motion perception) is crucial for developing a machine that can perceive, reason, and interact with real-world scenes based on 2D image inputs.

Given that the locations and motions of entities in physical scenes are represented in appropriate/efficient frames of reference, the problem of reasoning about inferred variables remains. Previous work has shown that “recognition” of dynamic events (e.g., discerning
whether an event appears natural) requires estimation of their physical likelihood given a Newtonian model (Sanborn, 2014; Sanborn et al., 2013). In object collision situations, this is achieved by comparing expected and observed velocity values (see Chapter 3). However, when more complex situations (e.g., ones including multiple objects or non-solid substances) are considered, the likelihood calculation becomes troublesome. Instead, people are generally asked to reason about the future status of a situation, which is mathematically expressed as the predictive probability of a candidate final state given an initial state distribution: \( P(S_f|S_i) \) (Bates et al., 2015; Battaglia et al., 2013; Hamrick et al., 2016; Kubricht et al., 2016). Chapter 5 demonstrated how this approach can be utilized to explain people’s predictions about a range of situations involving non-solid substances and collections of rigid objects. However, the differential equations governing substance dynamics are likely an implausible solution for the human brain: e.g., computational precision in implementing the differential equations may be outside of the capacity of neural encodings. Results from our ball approximation model (BAM) suggest that meaningful physical predictions can be made by coarsely approximating substances as relatively small numbers of rigid balls. By adding perceptual noise to the initial positions of the constituent balls—and applying rigid-body physical principles—we were able to demonstrate the viability of approximated simulation engines in predicting human expectations about substance dynamics. Importantly, recent deep learning approaches have demonstrated that rigid-body interactions can be learned from experience and applied to newly observed situations (Chang et al., 2016; Battaglia et al., 2016). A natural extension of the deep learning approach is to consider substance-related situations using an approximation algorithm similar to that of our BAM. If successful, the model could provide insight on how mental models explaining complex physical phenomena can be learned and encoded in neural circuitry. Human learning in intuitive physics is elaborated upon further in Section 7.1.

It is important to note that the reasoning component of intuitive physics extends beyond acquisition and application of physical models in perceived scenes. Instead, physical knowledge can be transferred to novel, seemingly unrelated situations unbeknownst to the observer (e.g., Day & Goldstone, 2011; Pedone et al., 2001). Chapter 6 provided evidence that the
depth with which a physical situation is understood depends on the working memory capacity of the observer. Moreover, quality of understanding is a significant predictor of whether an observer will transfer that knowledge to a novel target domain. Our spontaneous transfer results indicate that animations (i.e., dynamic event depictions) are particularly effective in demonstrating higher-level concepts that involve relations between entities over time. These concepts appear to extend beyond Newtonian combination rules and perceptual approximation, as previously discussed. Instead, physical events oftentimes demonstrate something greater—a schema, or theory which can be recalled and applied to new situations in the world. From this perspective, intuitive physics can be thought of as three levels of abstraction. At the bottom level is perception: e.g., where something is, how fast it is moving, how big it is, what it is made of. Computer vision is well-suited for extracting this type of information. Once low-level information is extracted, it proceeds to the next level which utilizes physical models to predict how situations evolve over time. The third and final stage develops a symbolic interpretation of the information in the preceding two levels. It is the bi-directional connection of these three levels which will constitute a significant direction of research in the field of intuitive physics during the years to come. Taken together, the work presented in this dissertation examines human cognition at each of the aforementioned levels and constitutes progress towards developing a computational system that can perceive, reason about, and interact with physical entities in the realizable world.

7.1 Developing a Physical Model of the World

The field of intuitive physics owes a great deal to early developmental work exploring infants’ expectations about the physical world (Baillargeon, 1987, 1994, 2002; Baillargeon, Spelke, & Wasserman, 1985; Kim & Spelke, 1999; Kellman & Spelke, 1983; Leslie, 1982; Leslie & Keeble, 1987; Spelke, Katz, Purcell, Ehrlich, & Breinlinger, 1994; Spelke, Kestenbaum, Simons, & Wein, 1995). Contrary to findings from adult studies during the same era (reviewed in Chapter 2), developmental results demonstrated that humans attain a wealth of knowledge about the physical world within the first six months of their lives. For example,
infants as young as four months old expect objects to move along continuous trajectories in space (Spelke et al., 1995), and infants as young as five months old understand that rigid objects support one another when stacked (Baillargeon, 1994). Under Spelke’s core knowledge thesis (Spelke & Kinzler, 2007), humans attain physical knowledge by building upon core physical principles, such as continuity and object permanence. In other words, implicit physical expectations develop in a piecemeal fashion—each subsequent component enriching the next. Moreover, informative physical variables are not represented with high fidelity from the start. For example, infants initially reason about support situations using a qualitative contact variable: if an object and surface are in contact with one another, the object will be supported, and vice versa. Later in their development, infants begin to understand that the amount of contact between an object and surface indicates whether the object will stay at rest or fall down. This characteristic of human learning is critically absent in current computational approaches to intuitive physics. Although models under the noisy Newton framework, for example, learn about the values of informative variables in dynamic situations, they are not required to learn which variables should be represented and attended to.

In the real world, there are an infinite number of candidate hypotheses regarding which objects are related to (i.e., physically interact with) one another. This is not a problem in the physical problems solved by deep learning (interaction) networks, since all of the objects in their 2D training displays can potentially interact with one another (Battaglia et al., 2016; Chang et al., 2016). For potentially large numbers of interacting objects in observed scenes, it would appear that deep learning methods would break down. Thus, a more elegant solution for considering candidate interaction hypotheses should be developed. It is my opinion that given observed physical scenes, a truly complete intuitive physics architecture should be able to discern (i) which physical variables to represent; (ii) how well to represent them and in what frame of reference; (iii) how they interact with one another; and (iv) how to combine interaction rules to build higher-level physical principles. Moreover, the architecture should achieve these sub-goals by passively observing 2D image data from a broad range of 3D scenes demonstrating a breadth of physical phenomena. To date, applications of deep
learning in intuitive physics are highly task-specific. Intuitive physical inference, however, can be probed by many different tasks with various types of entities (e.g., rigid objects or non-solid substances). It remains unclear how a task-specific model which learns physical interaction rules can generalize to novel domains. In short, the field of intuitive physics is far away from realizing the aforementioned capabilities. However, it is my hope that recent strides in machine learning and computational cognitive science will bring researchers closer to developing an intuitive physics architecture which can perceive, learn, and generalize understanding to novel situations in the world.

7.2 Do Explicit Conceptions Influence Physical Intuitions?

Chapter 5 showed how a problem which was originally posed at the explicit level (Schwartz & Black, 1999) can be accurately solved when reasoning implicitly through mental simulation. Whereas explicit reasoning draws on verbalizable (conceptual) knowledge, implicit reasoning tends to occur outside of awareness, yielding responses which generally cannot be explained. It is important to note that this explicit vs. implicit distinction extends across a breadth of tasks in the intuitive physics literature (Howe et al., 2014; Kaiser et al., 1986, 1992; Krist et al., 1993; McCloskey & Kohl, 1983; Schwartz & Black, 1999; Smith et al., 2013). The general consensus today is that explicit conceptions do not impact or interfere with implicit expectations. The main evidence for this stance is that people’s explicit performance does not appear to correlate with their implicit performance in classically studied tasks (Smith et al., 2013). However, it is my opinion that assuming dissociation between explicit and implicit reasoning in intuitive physics—across all problem-types—is reckless. This opinion is supported by a number of findings in the early intuitive physics literature demonstrating that explicit conceptions are sometimes (at least partially) linked to implicit performance on dynamic, realistic tasks.

For example, successfully solving the water-level task (WLT; introduced in Chapter 5; see Figure 5.1A) appears to rely on explicit understanding of the horizontality principle: i.e., the surface of water should remain horizontal regardless of the orientation of its container.
Howard (1978) found that each of his participants who could clearly state the horizontality principle prior to completing a dynamic version of the WLT gave correct responses. Although tilted water-level estimates are most likely due to a perceptual illusion stemming from usage of an object-centered frame of reference (McAfee & Proffitt, 1991), explicit understanding of horizontality appeared to prompt Howard’s participants to adopt an environment-centered frame of reference instead. This suggests that perceptual factors which lead to incorrect impressions of physical “naturalness” can be overcome by explicit awareness of a situation’s ground-truth characteristics. However, the corresponding finding requires further replication before a definitive conclusion can be drawn.

Studies have also shown that erroneous motion is implicitly perceived as unnatural when presented in a dynamic context (e.g., Kaiser, Proffitt, & Anderson, 1985) although people still rate ground-truth trajectories as unnatural in some cases. Unfortunately, Kaiser et al. (1985) did not report whether their participants’ explicit and implicit performance was correlated, although they did note that prompting observers to mentally simulate static depictions did not improve performance. An earlier study did, however, show that people sometimes plan and execute actions in accordance with their explicit misconceptions (McCloskey & Kohl, 1983). This is consistent with research demonstrating that children expect objects to move in the direction that they are pushed when playing video games, which is inconsistent with Newtonian physics (DiSessa, 1982). Although explicit understanding of the physical world does not appear to have a direct impact on implicit naturalness ratings, erroneous motor planning demonstrates that novel physical situations are (at least to some degree) considered at the explicit level during our everyday lives. Researchers today have argued that explicit and implicit understanding of the physical world is at least partially dissociated, although a suitable framework underlying the dissociation has yet to be developed (Howe et al., 2014). It is my hope that future research will (i) explore the role (if any) explicit knowledge has on people’s intuitions about the physical world, and (ii) work to develop a framework outlining the explicit and implicit reasoning systems interact in human intuitive physics.
7.3 Physical Judgments in Virtual Reality

As mentioned throughout this dissertation, a key shortcoming of early research in intuitive physics stems from the impoverished format in which dynamic situations were presented (e.g., Kaiser et al., 1986; Kaiser, Proffitt, & Anderson, 1985; McCloskey, 1983; McCloskey et al., 1980, 1983; Smith et al., 2013). Even situations presented in the form of dynamic 2D animations sometimes yielded human expectations that conflicted with results from studies utilizing realistic 3D materials (Flynn, 1994). Today, new forms of multimedia technology are becoming available for conducting human research. One notable technology which is beginning to gain traction in the field of psychology is that of virtual reality (VR; Wilson & Soranzo, 2015). In recent years, the quality and affordability of consumer-level VR devices has improved at an incredible rate, providing a valuable tool for constructing novel experimental environments to probe human cognition. Critically, such technologies were not readily available to researchers during the latter half of the previous century.

Compared to static 2D depictions of physical situations, VR provides an analog experience in a 3D artificial world from an embodied (egocentric) perspective. Environments constructed for human experiments have varied from those similar to real-world situations (Bruder & Steinicke, 2014; Knapp, 2016) to those which cannot be easily emulated in the natural world (Lopez et al., 2014; Schatzschneider et al., 2016; Schwartz & Black, 2016). Regarding intuitive physics, VR allows for the creation of environments whose characteristics differ from the real world in ways that preclude laboratory experimentation. In Chapter 4 I reported a study examining human judgments about projectile motion (including experiments adapted from previous work; Krist, 2000) under virtual gravity fields ranging from 50% to 150% of Earth’s gravity. The experiments shed light on the reasoning systems people utilize when gauging different aspects of projectile motion. Whereas estimates of flight duration were systematically biased towards expectations under Earth’s gravity field, predictions about the future location of a projectile were accurate regardless of the environment’s gravitational acceleration. By constructing an environment where the rate at which objects fall downward was systematically varied, we were able to show that location
prediction does not appear to involve “brute force” mental simulation like time estimation does.

In addition to modifying the quantity of known physical characteristics (e.g., the magnitude of gravitational acceleration) in virtual environments, my collaborators and I have also varied low-level spatiotemporal characteristics in intuitive displays (Lin, Zhu, Kubricht, Zhu, & Lu, 2017; Wang et al., 2018). In previous work, we examined how perceived causality (i.e., the degree to which an object appears to have launched another) in 3D collision events is influenced by modifications to the spatial locations of collision sounds in observed displays (Wang et al., 2018). This experiment examined the same type of launching event reported in Chapter 3 by replicating a previous study on audiovisual phenomenal causality (Guski & Troje, 2003). In this case, VR technology leverages the fact that the sound of two objects colliding cannot be easily “moved” to a different location in the real world: e.g., as occurs in the ventriloquist effect. While it has been shown that varying when the sound occurs influences perceived causality, an efficient method for varying where the sound occurs had not been reported. We found that people report weaker impressions of causality the further a collision sound is from the location of impact, which is a relatively trivial result. However, participants reported greater causal impressions when the sound emanated from where the objects were headed compared to where the objects had been. Moreover, when the collision sound emanated from the true location of impact, estimates of the sound’s location were marginally skewed in the direction of the collision (i.e., where the objects were headed). These results show that (i) classical results in 2D object collision judgments extend to realistic 3D situations; and (ii) perceived causality is inherently sensitive to novel spatiotemporal violations in dynamic events (i.e., violations that are not encountered in daily life). However, our results suggest that attention is not evenly distributed across space when observing collision events, which appears to influence the underlying ventriloquist effect in our VR experiments. Importantly, this is not consistent with previous research on multisensory perception (Bertelson, Vroomen, De Gelder, & Driver, 2000), which reports that deliberate visual attention does not influence the underlying effect. Future work should further utilize VR technology to systematically examine sensory integration in object collision
situations (with spatially perturbed auditory collision indicators) to test the replicability of our finding.

A final result from our research points to an important shortcoming of VR environments regarding human interaction with virtual entities. Namely, when striking an object to propel it forwards, the object appears weightless due to the absence of haptic feedback following the interaction (see Chapter 4, Experiment 1). Because participants implicitly gauged objects as being lighter than they actually were, they imparted upon them a smaller force than needed when propelling them to target locations. In intuitive physics research, the general goal is to assess human intuitions about the attributes, dynamics, and relations between entities in novel environments. When intuitions follow from visual observation, VR environments are a reliable tool for probing human expectations. However, when physical information is attained through direct interaction, the brain forms expectations about haptic signals which are absent in the virtual world. Even the experience of one's own weight arguably interferes with the perception of unrealistic gravitational acceleration, as mentioned above. A key direction for future work in human VR experiments is developing apparatuses which can emulate haptic feedback to provide a true analog experience to the physical world. One promising recent approach is to leverage the dominance of vision by using a single physical prop to “generate” haptic information for multiple virtual objects (Azmandian, Hancock, Benko, Ofek, & Wilson, 2016). It is my hope that future work examining human cognition in VR environments will consider the directions reported herein.
APPENDIX A

Background and Outstanding Questions (Chapter 2)

A.1 Erroneous Theories of Motion

The motions of objects and substances in the world are accurately described by the principles of Newtonian physics. Newton’s three laws state that (i) an object at rest will stay at rest, and an object in motion will stay in motion, unless acted upon by an external force, (ii) objects accelerate when an external force is applied to them, and (iii) for every action, there is an equal and opposite reaction. Therefore, when objects fall, a gravitational force accelerates them downwards, and when a force is applied to an object in a given direction, the object retains the motion it had before the force was exerted.

By contrast, medieval impetus theory states that (i) the act of setting an object in motion imparts upon it an ‘impetus’ force that is used to maintain its motion, and (ii) the impetus force of a moving object gradually dissipates over time (McCloskey, 1983). Thus, an object thrown in the air falls down because its vertical impetus dissipates, and an object will continue along a curved trajectory after it is released from a pendulum because of a (dissipating) curvilinear impetus force.

The principles of Aristotelian physics state that an object will move in the direction that it is pushed (DiSessa, 1982). In other words, if a horizontally moving object is given a vertical push (e.g., a gravitational force downwards), it will immediately move straight downwards and lose its horizontal component of velocity.
A.2 Early Research on Intuitive Physics

Early research in the field provides several examples of situations in which humans demonstrate common misconceptions about how objects in the environment behave. The most extensive line of research involves human judgments about object collisions, specifically about which object appears heavier based on an animated depiction of the collision event. Interestingly, people have a tendency to believe that the object that moves with a greater initial velocity is heavier (Todd & Warren Jr., 1982). This bias has recently been reinterpreted within the probabilistic simulation framework (Sanborn, 2014; Sanborn et al., 2013).

Another major line of research examined human knowledge about projectile motion. In these tasks, participants viewed a static depiction of an object moving in a given situation and were asked to draw the trajectory the object would follow as time progressed. In the falling object problem, people commonly responded that an object dropped from a moving body will follow a straight path downwards (McCloskey et al., 1983). People’s predictions about the trajectory of an object after being released from a pendulum were also inconsistent with Newtonian principles (Cohen, 1981), but a probabilistic simulation model has achieved success predicting where the object is expected to land (Smith et al., 2013). People also commonly predict that an object exiting a curved tube will follow a nonlinear trajectory upon exiting (McCloskey & Kohl, 1983). Although people’s trajectory predictions in these problems are inconsistent with normative physical principles, they correctly judge anomalous trajectories as appearing unnatural when presented in an animated format (Kaiser et al., 1992). Currently, no computational account explains people’s erroneous misconceptions in explicit trajectory prediction problems.

Work on mechanical reasoning supports the role of mental simulation in intuitive physics. In the pulley problem, people take longer to respond for pulleys farther from the beginning of the causal sequence (i.e., the rope being pulled; Hegarty, 2004). This finding suggests that humans build analog representations of physical situations that carry spatial information, as opposed to reasoning according to a simple rule (e.g., alternating pulleys rotate in opposite directions). The pervasiveness of mental simulation is further reinforced by results from
the water-pouring task (Schwartz & Black, 1999; see Figure 2.1f). In this task, participants were successful when forming predictions using mental simulation but performed much worse when solving the problem using explicit reasoning. A probabilistic simulation model has recently accounted for people’s mental simulation performance in a modified water-pouring task (Kubricht et al., 2016). Taken together, these problems provide several examples in which people’s explicit reasoning about the physical world is biased, and suggest experimental factors that lead to accurate predictions and judgments.

A.3 Object Collision Physics

In the case of collision events, two objects move towards one another, collide, and then move away from one another. The movement of each object both before and after the collision is represented by its velocity, which specifies the speed and direction of each object. While velocity can be perceived, the mass of each object (i.e., how heavy each object is) and the amount of restitution (i.e., how much energy is given back’ in the collision) cannot.

The momentum of an object is defined by the product of its mass and velocity. The principle of conservation of momentum states that, in a closed system, the sum of the momentum of objects before a collision is equal to the sum of their momentum after (i.e., momentum is conserved). By rearranging the conservation of momentum expression for two-body collisions, the relative mass of two objects can be expressed purely in terms of the initial and final velocities of the objects (Runeson, 1983). Thus, if humans can perceive the velocity of objects without error and reason in accord with momentum conservation, their relative mass judgments should be invariant to the amount of restitution in the collision. This is termed the direct perception model. Under this view, people should be equally accurate when reasoning about the relative masses of two billiard balls (high restitution) as they are when reasoning about those of two rubber balls (low restitution). However, humans are increasingly inaccurate when reasoning about relative mass in low-restitution collisions, thus deviating from what would be expected given Newtonian physical principles.
A.4 Trends

- People demonstrate common misconceptions when asked to make explicit reasoning judgments about physical systems on pencil-and-paper tasks. However, their performance improves when the problem is accompanied with dynamic and contextual information.

- Recent research has shown that our implicit judgments about physical situations based on rich dynamic displays are consistent with probabilistic mental simulation governed by normative physical principles, but are subject to biases evoked by perceptual uncertainty and prior beliefs about physical variables.

- Recent behavioral evidence suggests that people utilize a cognitive intuitive physics engine to reason about a broad range of physical situations.

- Researchers have begun to explore how to integrate probabilistic mental simulation with deep-learning models to extract higher-level physical knowledge from static and dynamic visual inputs.

A.5 Outstanding Questions

- How proficient are people at reasoning about the dynamics of relatively unfamiliar non-rigid substances (e.g., sand, honey), and what prior knowledge do people adopt about the attributes of unfamiliar substances?

- What perceptual characteristics of intuitive physics problems are necessary to enable spatial representation of physical variables and subsequent mental simulation?

- What is the role of dynamic uncertainty (i.e., uncertainty in people’s internal model of physical dynamics) in mental simulation? What factors cause changes in the amount of uncertainty, and how can stochastic noise at different levels of processing be implemented into numerical physics models?
• How do humans acquire the ability to perform mental simulation, and selectively control the use of mental simulation in different situations? What computational constraints and processing constraints are adopted when reasoning about physical situations?

• How are intuitive mental models of Newtonian physics implemented in the neural circuitry of the brain?

• To what extent can learning-based models (such as deep-learning models and other neural networks) emulate physical knowledge and explain intuitive physics?
APPENDIX B

Model Details (Chapter 3)

B.1 Noisy Newton Model for Mass Inference

The noisy Newton model for mass judgment uses Bayes’ rule to calculate a posterior distribution of collision attributes, \( A \), given noisy observable information \( O \):

\[
P(A|O) = \frac{P(O|A)P(A)}{P(O)}, \tag{B.1}
\]

where \( P(A) \) represents prior knowledge people have about hidden attributes in collision events. Those attributes are mass, \( m_A \) and \( m_B \), and restitution, \( e \), which is a constant between 0 and 1 that represents the elasticity in a collision. The model assumes that all restitution values are equally likely and objects are more likely to be light than heavy:

\[ e \sim \text{Uniform}(0,1); \ m_A, m_B \sim \text{Exponential}(1). \]

The \( P(O|A) \) term indicates the likelihood of observed velocities \( O = u_A, u_B, v_A, v_B \) given a potential set of attributes \( A = e, m_A, m_B \). Here, \( u_A \) and \( u_B \) are the pre-collision velocities of Objects A and B, and \( v_A \) and \( v_B \) are the post-collision velocities. Post-collision velocities are calculated based on the pre-collision velocities, the object masses, and the collision’s restitution coefficient via the following equations:

\[
v_A = \frac{m_A u_A + m_B (u_B + e (u_B - u_A))}{m_A + m_B} \tag{B.2}
\]

\[
v_B = \frac{m_B u_B + m_A (u_A + e (u_A - u_B))}{m_A + m_B} \tag{B.3}
\]

A noisy observation model then links true, hidden variables \( \bar{O} \) with observed variables \( O \).
such that their difference \( \epsilon \) is normally distributed in logarithmic space: \( \epsilon \sim \text{Gaussian}(0, k^2) \). Given a weighted logarithmic transformation function \( f(x) = \text{sign}(x)\log(w|x| + 1) \), the difference between observed and true observations is expressed as \( \epsilon = f(O) - f(\bar{O}) \).

With the noisy observation model, the \( P(O|A) \) in Equation B.1 can be expanded to include both \( O \) and \( \bar{O} \):

\[
P(O|A) = \int_{\bar{O}'} P(O|\bar{O}') P(\bar{O}'|A)d\bar{O}',
\]

where the \( P(\bar{O}|A) \) term is further separated into initial and final velocities: i.e., \( P(\bar{O}|A) = P(\bar{v}_A, \bar{v}_B|A)P(\bar{u}_A, \bar{u}_B) \). Note that pre-collision velocity does not depend on the collision attributes. Instead, values for \( \bar{u}_A \) and \( \bar{u}_B \) are drawn from the slow motion prior such that \( \bar{u}_A, \bar{u}_B \sim \text{Gaussian}(0, \sigma^2) \). Post-collision velocities are then calculated from Equations B.2 and B.3.

### B.2 Noisy Newton Model for Causal Inference

The noisy Newton model can also be used to predict the marginal probability of a causal relationship, \( C \), given noisy observable information, \( O \):

\[
P(C|O) = \frac{P(O|C)P(C)}{P(O|C)P(C) + P(O|NC)P(NC)},
\]

The \( P(O|C) \) term in Equation B.5 can be expanded to the following integral:

\[
P(O|C) = \int_{\bar{O}', A'} P(O|\bar{O}') P(\bar{O}'|A', C) P(A')d\bar{O}'dA'.
\]

Note that temporal delay, \( t \), can be included as an observable variable with log-normal uncertainty and a delta function prior centered at 0 msec: \( P(\bar{t}) = \delta(\bar{t}) \). For the \( P(O|NC) \) term, Sanborn et al. (2013) set this value as a free parameter in their model. The authors also made the assumption that causal and noncausal models were equally likely: i.e., \( P(C) = P(NC) \).
APPENDIX C

Simulator Details (Chapter 5)

C.1 MPM Simulator Details

The governing partial differential equations utilize the principles of conservation of mass and momentum:

\[
\frac{D\rho}{Dt} + \rho \nabla \cdot \mathbf{v} = 0, \quad \frac{D\mathbf{v}}{Dt} = \nabla \cdot \mathbf{\sigma} + \rho \mathbf{g},
\]

where \(\mathbf{\sigma}\) is the stress imparted on a particle, \(\mathbf{g}\) is the gravitational acceleration, and \(\frac{D}{Dt}\) is the material derivative with respect to time. The equations are discretized spatially and temporally with a collection of Lagrangian particles (or material points) and a background Eulerian grid. The material type of the simulated substances is naturally specified from the constitutive model, which defines how a material exerts internal stress (or forces) as a result of deformation.

Rigid balls are simulated as highly stiff elastic objects with the neo-Hookean hyperelasticity model, described through the elastic energy density function

\[
\Psi(\mathbf{F}) = \frac{\mu}{2} (tr(\mathbf{F}^T \mathbf{F}) - d) - \mu \log(J) + \frac{\lambda}{2} \log^2(J),
\]

where \(d\) is the dimension (2 or 3), \(\mathbf{F}\) is the deformation gradient (i.e., the gradient of the deformation from undeformed space to deformed space), \(J\) is the determinant of \(\mathbf{F}\), and \(\mu\) and \(\lambda\) are Lamé parameters that describe the material’s stiffness.

Liquid is modeled as a nearly incompressible fluid, with its state governed by the Tait equation (Batchelor, 2000):

\[
p = k \left[ \left( \frac{\rho_0}{\rho} \right)^\gamma - 1 \right],
\]
where $p$ is the pressure, $\rho$ and $\rho_0$ are the current and original densities of the particles, $\gamma = 7$ for water, and $k$ is the bulk modulus (i.e., how incompressible the fluid is). Through this Equation-of-State (EOS), the stress inside a non-viscous fluid is given by $\sigma = -p\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. We further adopt the Affine Particle-In-Cell method (APIC; Jiang et al., 2015) to greatly reduce numerical error and artificial damping. This enables us to simulate fluids with better accuracy compared to alternative computer graphics methods.

The motion of dry sand is largely determined by the frictional contact between grains. In the theory of elastoplasticity, the modeling of large deformation (e.g., frictional contact) can be based on a constitutive law that follows the Mohr-Coulomb friction theory. Following (Klár et al., 2016), we simulate dry sand based on the Saint Venant Kirchhoff (StVK) elasticity model combined with a Drucker-Prager non-associated flow rule. Plasticity models the material response as a constraint projection problem, where the feasible region (or yield surface) of the final material stress is restricted to be inside

$$tr(\sigma)c_F + \sigma - \frac{tr(\sigma)}{d} F \leq 0,$$

where $d$ is the dimension and $c_F$ is the coefficient of internal friction between sand grains. The stress (and thus deformation gradient) of each sand particle is projected onto the yield surface so as to satisfy the second law of thermodynamics.
APPENDIX D

Experiment Materials (Chapter 6)

D.1 Source-Understanding Questions

Multiple-Choice Questions:

• What level of damage did the cannon ball(s) cause to the friendly barrier in Scenario [1–4]?

• What level of damage did the cannon ball(s) cause to the enemy octagon in Scenario [1–4]?

Free-Response Questions:

• Why was the cannon [were the cannons] unable to defeat the enemy octagon in Scenario [1–3]?

• Why were the cannons able to defeat the enemy octagon in Scenario [4]?

D.2 Radiation Problem

Story: Suppose you are a doctor faced with a patient who has a malignant tumor in his stomach. It is impossible to operate on the patient, but unless the tumor is destroyed, the patient will die. There is a kind of ray that can be used to destroy the tumor. If the rays reach the tumor all at once at a sufficiently high intensity, the tumor will be destroyed. Unfortunately, at this intensity the healthy tissue that the rays pass through on the way to the tumor will also be destroyed. At lower intensities, the rays are harmless to healthy
tissue, but they will not affect the tumor either. What type of procedure might be used to destroy the tumor with the rays, and at the same time avoid destroying the healthy tissue?

**Question Before Hint:** Please write down as many possible solutions you can think of, without worrying about whether solutions would really work. No medical expertise is required.

**Question After Hint:** Think back to the scenarios you observed earlier and write down any solution to the radiation problem they may suggest. It’s okay to repeat a solution you gave earlier.

### D.3 Scoring Criteria

**Scenario 1:**

- The small cannonballs inflicted no damage to the octagon.
- There was only one cannon firing cannonballs at the enemy octagon as opposed to multiple cannons.
- The small cannonballs inflicted no damage to the friendly barrier.

**Scenario 2:**

- The large cannonballs inflicted minor/some damage to the octagon.
- There was only one cannon firing cannonballs at the enemy octagon as opposed to multiple cannons.
- The large cannonballs inflicted major/substantial damage to the friendly barrier.

**Scenario 3:**

- The multiple large cannonballs inflicted substantial/major damage to the octagon (by *converging* onto their target *simultaneously*).
• There were multiple cannons firing cannonballs at the enemy octagon as opposed to a single cannon.

• The large cannonballs inflicted substantial/major damage to the friendly barrier.

Scenario 4:

• The multiple small cannonballs inflicted some/moderate damage to the octagon (by converging onto their target simultaneously).

• There were multiple cannons firing cannonballs at the enemy octagon as opposed to a single cannon.

• The small cannonballs inflicted no damage to the friendly barrier.
References


dict liquid dynamics using probabilistic simulation. In *Proceedings of the 37th annual conference of the cognitive science society*.


York: Oxford University Press.


154


Five-month-old infants have general knowledge of how nonsolid substances behave and interact. *Psychological Science, 27*(2), 244–256. doi: 10.1177/0956797615617897


machines that learn and think like people. Behavioral and Brain Sciences, 40. doi: 10.1017/S0140525X16001837


Lopez, T., Bouville, R., Loup-Escande, E., Nouviale, F., Gouranton, V., & Arnaldi, B.


Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science, 10*(1),
89–96. doi: 10.1111/j.1467-7687.2007.00569.x


conceptions of motion based on a naive theory or on prototypes? *Cognition*, 29(3), 251–275. doi: 10.1016/0010-0277(88)90026-1


