Augmented Reality Water-Level Task

Romina Abadi

A Thesis submitted to the Faculty of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science.

Graduate Program in Computer Science.

York University, Toronto, Ontario.

January 2023

Abstract

The"Water Level Task" asks participants to draw the water level in a tilted container. Studies showed that many adults have difficulty with the task. Up to now, most studies of the Water-Level Task have been based on schematic 2D presentations. Research on intuitive physics has found that people make fewer errors when judging physical phenomena in a more natural context.

Our study aimed to take an intuitive physics approach to analyze human performance on the Water-Level Task. We created an augmented reality (AR) version of the task (AR-WLT) to determine if the misconception about water orientation happens in a more natural environment. Moreover, a traditional online WLT was created to recruit low and high-scoring participants. We implemented a water-in-container effect using Unity shaders. In the AR WLT task, participants interacted with two containers half full of water in a Hololens2 AR display and were asked to determine which looked more natural. In at least one of the two simulations, the water surface did not remain horizontal when the container was tilted. Almost all low-scoring participants drew lines parallel to the bottom of the container in the online WLT task. Thirty-five low and high-scoring individuals participated in the AR-WLT. Our AR-WLT results showed that people prone to errors in the 2D version of the task were more likely to make errors in the Augmented Reality version. However, they did not choose simulations close to their 2D drawings. We also found that all participants were insensitive to minor tilts of the water surface.

Dedication

To my sister Tina, my number one source of motivation and strong support during the quarantine days.

Declaration

I declare that this thesis has been composed by myself and the results are my original work.

Romina Abadi

Jan 2023

Table of Contents

List of Tables

List of Figures

Chapter 1

The Water-Level Task

1.1 Introduction

In its standard form, the Water-Level Task (WLT) is a spatial cognition task in which the participants see a two-dimensional outline drawing of a vertically oriented bottle containing water and are asked to draw the edge of water surface in pictures of the same container tilted to different degrees (Figure 1.1).

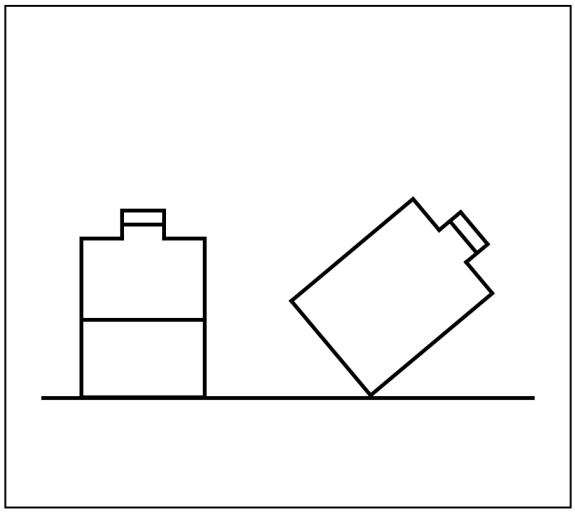


Figure 1.1: Water-Level Task example based on (Vasta & Liben, 1996). Participants are asked to determine the water level in the tilted containers.

Piaget and Inhedler originally developed the WLT to examine spatial cognition development in children of different ages (Piaget & Inhelder 1956, cited from Vasta & Liben 1996). They observed that young children did not draw horizontal waterlines even after seeing an actual tilted container half full of water. They proposed that as the spatial cognition abilities in children develop, they form a horizontal and vertical frame of reference with which they perceive the world. Therefore, children can understand that the water level remains horizontal and thus depict it correctly by about age nine (Kalichman 1988, cited from Piaget & Inhelder 1956). However, Rebelsky (1964) reported that many of his university students struggled with the task. Since then, many studies have focused on adult performance on the WLT. It has been reported that approximately 40% of adults draw lines that deviate more than 5° from the horizontal (Proffitt & Kaiser, 2006).

Although the Water-Level Task was initially created to evaluate spatial cognition, it can also be viewed from an "Intuitive Physics" perspective. The WLT is one of many cases where people are inaccurate in illustrating the physical world. Intuitive Physics research focuses on studying instinctive human understanding of physical principles, especially ones affecting interactions with the world in everyday life (J. R. Kubricht et al., 2017; Proffitt & Kaiser, 2006). Research has shown that although people are accurate and efficient in their everyday interactions with the physical world, their explicit judgement of physical events (e.g. articulating physical principles or predicting the state of the world in abstract drawings) is prone to errors (Smith et al., 2013). The Water-Level Task can be viewed as an Intuitive Physics study as it requires an understanding of gravity and how it affects the orientation of the liquid's surface.

This chapter summarizes the previous findings on adults' performance in the WLT and the related findings and studies on Intuitive Physics.

1.2 The Water-Level Task

Study of adult performance on the Water-Level Task has shown that a significant proportion of adults do not draw horizontal water level in tilted containers (Kalichman, 1988; Proffitt & Kaiser, 2006; Rebelsky, 1964; Vasta & Liben, 1996). In different studies, different ratios of people have been reported to have difficulty with the task. The reported number of people who draw tilted lines depends on the details of the studies, such as population, task difficulty, and the performance measure.

Researchers have mainly adopted two approaches to measuring performance on the traditional Water-Level Task. Some researchers have used an individual's rate of successful trials (the depicted line being tilted less than an acceptance threshold from horizontal) out of the total trials. In contrast, others have used the average absolute tilt from the horizontal for all trials as one's score and used a threshold to determine whether the participant was successful (Vasta et al., 1993). Most researchers have used the acceptance threshold of 5°; however, some researchers have adopted an acceptance threshold of 10° to allow more room for errors caused by inaccuracy in drawing horizontal lines (e.g. Barhorst-Cates et al. 2020).

It has been reported that the shape and tilt of the containers affect the task difficulty. McAfee & Proffitt (1991) observed that symmetrical containers, such as round containers or bowls, made the tasks easier. Also, (Vasta et al., 1996) reported that more participants were accurate when containers' tilts from upright were smaller.

A recurring observation in WLT studies was that, on average, men were more successful than women. For example, Sholl & Liben (1995) reported that 17% of males and 39% of females among 900 undergraduate students failed to draw a horizontal water edge in at least seven out of eight tilted containers. Also, in a revised Water-Level Task, Robert & Morin (1993) asked participants to determine the water level in a watering can's spout. Thirty-eight percent of men and 61% of women did not draw a horizontal waterline. The

observed difference between female and male performance led to further research focusing on explaining the factors causing the gender gap, and trying to eliminate it (Kalichman, 1988; Vasta & Liben, 1996). Another line of research has focused on different perceptual biases and cognitive strategies that affect one's performance on the WLT (Kalichman, 1988; Proffitt & Kaiser, 2006; Vasta & Liben, 1996). This section summarizes the previous findings and outlines the variants of the Water-Level Task researchers have previously used.

1.2.1 Gender Gap Studies

Researchers have identified many factors that could explain the gender gap. The examined biological factors include handedness (Annett, 1994; Casey, 1996; Robert & Harel, 1996), gonadal hormone levels (Collaer & Hines, 1995), X-linked genes (Thomas & Jamison, 1981), and brain activity during the WLT (Wu et al., 2017). Studies have also explored the effect of sociological factors such as profession (Hecht & Proffitt, 1995; Vasta et al., 1997); field of study (Robert & Harel, 1996); and activities that develop spatial skills (Baenninger & Newcombe, 1995; Casey, 1996; Robert & Harel, 1996). Although researchers found correlation between the above mentioned factors and WLT performance, neither of the factors was sufficient to completely describe the difference between male and female performance.

In addition to explaining the gender gap, some researchers focused on strategies to eliminate it. For example, Vasta et al. (1996) observed that a self-discovery training procedure improved females' performance significantly. Men and women who completed the self-discovery training performed equally on the tasks.

Although the gender gap has been observed in most studies, some did not find a significant gender effect. As an example, in Barhorst-Cates et al. (2020)'s results, neither gender nor age significantly correlated with the performance of 8 to 11 years old children, suggesting the gender gap emerges at older ages. Moreover, in a few studies on adults, gender was not found

to be a significant factor. For example, McAfee & Proffitt (1991) reported that although men performed better in the traditional task, the performance of male and female participants was not significantly different when they were asked to make judgements about anomalous water orientations in photos. Also, Kenyon (1984) did not find a significant difference in the performance of 15 male and 15 female college students on a pen-and-paper variation of the WLT. A circle with clock markings was drawn outside the containers in their version of the task.

In general, although gender significantly correlates with participants' average performance on the WLT, it does not predict one's success. In other words, although the average scores of men and women are significantly different, individual women and men who perform accurately are equally accurate and unsuccessful individuals make similar errors regardless of their gender. Thus, some researchers have found it more instructive to focus on individual differences between successful and unsuccessful individuals rather than males and females.

1.2.2 Training

Research has shown that indirect training (e.g. mathematics course) and direct training (task-specific) improve one's performance in spatial skills tasks (Baenninger & Newcombe, 1995). To eliminate the gender gap, researchers designed implicit and explicit training procedures for WLT and evaluated participants before and after training.

Liben & Golbeck (1984) observed that showing one image of tilted container half full of water did not affect participants' performance. Explicitly telling participants the horizontality rule before the task improved their performance; however, it did not eliminate the gender gap. This evidence suggests that knowledge of the principle was not the only factor causing the differences.

Thomas et al. (1973) used two interactive training procedures. They used a 3D apparatus in which two bottles were mounted and tilted to the same degree. One bottle was half full of coloured water (the model bottle), and one bottle was empty. At the beginning of each trial, the two bottles were tilted to the same degree. They designed two learning procedures in which the participants were asked to adjust a pretend water line in the empty container to represent the liquid and could look at the model. However, only 3 out of 24 female participants who did not know the horizontality principle prior to the experiment could correctly adjust the pretend waterline without looking at the model at the end of 24 training rounds. Thomas et al. (1973) observed that the training procedure did not affect the participant's ability to state the horizontality principle.

Vasta et al. (1996) designed a self-explanatory learning procedure in which the WLTs were presented to the participants from easy to hard. At the beginning of the procedure, participants solved tasks in which the water level in a tilted container was given, and they were asked to determine the water level in an upright container. As the procedure continued, the tasks got progressively more complex. In other words, tilted containers with smaller tilts and simpler shapes were included earlier than more complex ones. Vasta et al. (1996) observed that the self-explanatory learning procedure eliminated the gender gap on the WLT. Moreover, women who did the learning procedure were significantly more likely to identify the horizontality principle than the control group who had completed the conventional WLT tasks. The training procedure did not significantly affect men's performance or ability to identify the horizontality principle. However, more men than women stated the horizontality principle among the trained participants. Moreover, the training procedure did not affect participants' performance on a slightly different task in which subjects were asked to draw a hanging object in a tilted van. Vasta et al. (1996) concluded that the gender gap on the WLT has an experiential component, meaning that women's childhood activities may involve less spatial training, which could contribute to their less accurate performance. The study

suggests that appropriate training could potentially eliminate this gap, whether it is caused by biological or external factors.

Based on the two studies discussed (Vasta et al. (1996) and Thomas et al. (1973)), abstract representations, such as 2D drawings, are more effective for training individuals on the WLT compared to using a real bottle. This is likely because the abstract setting simplifies the task by eliminating irrelevant information, such as the movement and color of the liquid, and facilitates the conceptualization of the horizontality principle. Also, it might be more difficult to determine water surface is horizontal in a real 3D bottle than in a 2D drawing. These findings suggest that the effectiveness of training may depend on the type of representation used and the nature of the task.

1.2.3 Expertise and Education

It has been observed that the field of study correlates with the performance on the WLT, with engineering students outperforming students from other disciplines (Robert & Harel, 1996). Some researchers believe that mathematics and physics training affects one's performance on spatial cognition tasks (Baenninger & Newcombe, 1995); however, the better performance of engineering students could also be attributed to their knowledge of the underlying physics principles. Some researchers have found that being able to explicitly articulate that the water always remains horizontal is a good predictor of the performance on the WLT (Howard, 1978; Thomas & Jamison, 1975). However, other researchers observed that the explicit knowledge of the principle does not always result in a better performance on WLT and vice versa (McAfee & Proffitt, 1991; Myer & Hensley, 1984; Robert & Harel, 1996; Vasta et al., 1996). For example, Myer & Hensley (1984) observed that among 85 university students, 37% of high-scoring participants (participants who drew accurate horizontal lines in the WLT task) did not know the horizontality principle explicitly, while 28% of low-scoring participants

(participants whose drawn lines were not within 5° of horizontal) stated that the water remains horizontal.

One interesting observation, reported by Hecht & Proffitt (1995), was that among different occupations (bartenders and waitpersons, students, homemakers, and truck drivers), bartenders were less accurate than the others despite interacting with containers of liquid more frequently. Hecht & Proffitt (1995) suggested bartenders focus on the container rather than the environment while interacting with it; thus, they tend to imagine the water level relative to the container rather than the environment. The choice of frame of reference is discussed in more detail in section 1.2.4.

1.2.4 Cognitive and Perceptual Differences

Some researchers focused on perceptual biases and cognitive strategies that affected performance on the WLT. In such studies, researchers typically choose balanced groups of high-scoring and low-scoring men and women based on their performance on the conventional WLT and analyze the individual differences (Kalichman, 1988).

Tilt Illusion and frame of reference

The existence of a tilted frame affects one's perception of the horizontal and vertical axes. Research has shown that in a tilted frame, the lines perceived as horizontal and vertical are tilted towards the same direction as the frame's tilt (Coren & Hoy, 1986; Goodenough et al., 1979) (Figure 1.2). The extent to which an individual is affected by this Tilt Illusion varies, and research showed that on average, women are more sensitive to Tilt Illusion than men (Sholl & Liben, 1995; Vasta et al., 1993; Vaught, 1965). The Tilt Illusion reaches its peak when the frame is tilted around 20°. Individuals experiencing tilt illusion perceive a line that is tilted to the same direction as the frame as horizontal. The maximum tilt angle

for a line that is perceived as horizontal is on average 10° (Goodenough et al., 1979; Sholl & Liben, 1995).

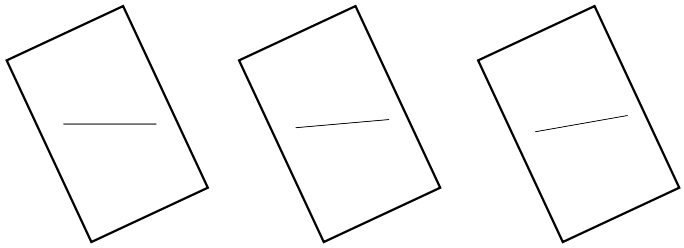


Figure 1.2: The Tilt Illusion effect. The lines in the middle and right pictures are tilted 5° and 10°, respectively, towards the frame tilt (counterclockwise). A person experiencing Tilt Illusion perceives the middle or right line as horizontal. As a result, a horizontal line (the left image) is perceived to be tilted in the opposite direction of the frame tilt (clockwise).

Research has shown that one's performance on the WLT correlates with their sensitivity to Tilt Illusion (Sholl & Liben, 1995; Vasta et al., 1993). The Tilt Illusion can affect one's performance on the WLT in two ways. First, individuals who know that water remains horizontal might not draw a horizontal line accurately because their drawings are affected by the presence of a tilted container (similar to the effect of a tilted frame in Tilt Illusion). Second, the misconception of horizontal in everyday life might prevent them from acquiring the relevant knowledge.

Vasta et al. (1993) found a correlation between the explicit knowledge of the horizontality of water and one's ability to draw a horizontal line, suggesting that spatial abilities might affect their acquisition of the physics principle. Moreover, Sholl & Liben (1995) observed that the low-scoring participants (participants who scored low in the traditional task) were less accurate than the high-scoring participants (participants who performed accurately on the traditional task) in identifying horizontal edges of liquid in tilted containers in video

segments. They suggested that the poor perception of horizontal in real life had affected the low-scoring participants' knowledge of the principle. However, it is possible that the high-scoring participants relied on their knowledge of the principle rather than perception.

Howard (1978) created stereoscopic videos and images of a container of Coca-Cola in front of a plywood wall with a shelf attached to it as the horizontality cue. By rotating the wall and camera, they created anomalous videos in which when the container tilted to 45 degrees (from the perceived upright position), the liquid did not remain horizontal and tilted to different degrees from horizontal (-30 to 50) (Figure 1.3). The stimuli consisted of stereoscopic videos and images with anomalous and correct water orientation. For each stimulus, the participants were asked to determine if the liquid was "the way normal liquid would be". Their results showed that individuals who knew the horizontality principle could perfectly identify anomalous videos and images. However, participants who did not know the horizontality principle accepted a tilt from -20 to $+10$ as natural. After being told the horizontality principle, 8 out of 12 participants perfectly identified the anomalous videos showing that their perception of horizontal was accurate. Moreover, one participant remembered the principle halfway through the task and scored perfectly afterwards. Thus, Howard (1978) proposed that the perception of horizontality does not precede knowledge acquisition.

Some researchers consider that the correlation between the sensitivity to tilt illusion and performance on WLT is not a cause-and-effect relationship. They suggest that one's preferred frame of reference affects their performance in both tasks (McAfee & Proffitt, 1991). In other words, susceptible individuals focus on the frame or the container rather than the environmental cues.

McAfee & Proffitt (1991) observed that implicitly redirecting participants' attention to the environment rather than the container improved their performance. To redirect the participants' attention, they designed a variant of WLT in which they created images of

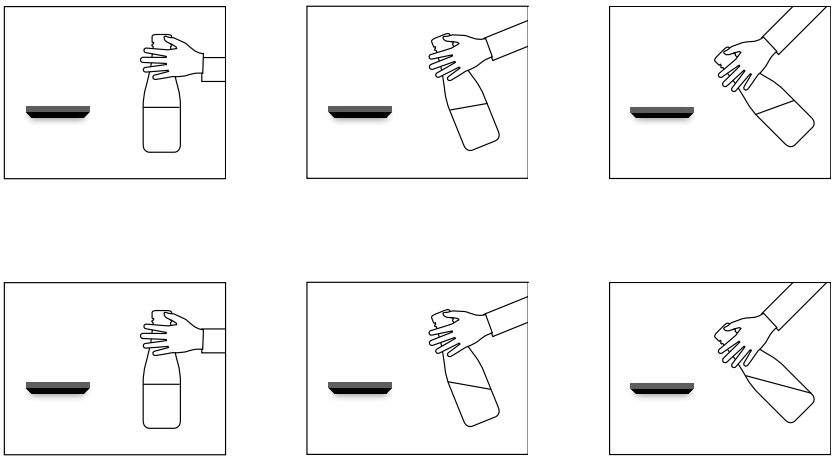


Figure 1.3: An illustration of anomalous liquid orientation in stereoscopic video frames used by Howard (1978). A shelf (depicted as a black rectangle) was present in the frames as a horizontal cue. The figure depicts two conditions in which the liquid tilts $+20°$ (top) and $-20°$ (bottom) when the bottle is tilted from $0°$ to $45°$ (left to right). Howard (1978) asked participants to judge whether the frames looked natural. This figure is created based on (Howard, 1978).

tilted containers with a solid background and cropped to a circular image. Then they asked subjects to orient the images, so they looked upright. The low-scoring participants' performance improved significantly compared to their performance on the pen-and-paper task; however, they were still less successful than the high-scoring participants.

McAfee & Proffitt (1991) observed a general tendency to accept liquid orientations tilted in the same direction as the container. In a pen and paper WLT, 8% of high-scoring and 89% of low-scoring participants were more likely to draw lines tilted in the same direction as the container, and the remaining participants were indifferent (equally likely to draw lines in either direction). Moreover, McAfee & Proffitt (1991) digitally created valid and anomalous images of a tilted container half full of coloured water over a real background scene with horizontal cues. In the anomalous images, the liquid was tilted $10°$ in the same

or the opposite direction of the container. Each resulting picture was cut round and tilted 20° to either side, so the only horizontality cue in the images was the background picture (Figure 1.4). They asked participants to judge if the picture was natural. Both low-scoring and high-scoring participants spent more time on the anomalous pictures in which the liquid was tilted in the same direction as the container (10°) and made the most errors in that setting. The condition that the liquid tilted abnormally in the opposite direction of the container was the easiest for participants to identify. This conclusion contrasts Howard (1978)'s observation that abnormal water tilts opposite to the container tilt direction looked natural for up to 20°, and for the tilts in the same direction, the liquid looked natural for up to 10° (i.e. participants were more sensitive to abnormal tilts in the same direction as the container tilt). The reason that different anomalies seemed natural in the photos and stereoscopic videos needs to be clarified. It should be noted that the same participants have not been tested in both settings, and the difference can be attributed to a difference in the populations. However, it can also be caused by the difference between the two tasks. More specifically, Howard (1978) used dynamic, stereoscopic depictions while McAfee & Proffitt (1991)'s photos were two-dimensional and static.

Spatial Cognition Style and Abilities

Studies found a strong correlation between performance on WLT and other spatial cognition tasks. Example of tasks include mental rotation (Ekstrom & Harman, 1976; Shepard & Metzler, 1971), embedded figures (Witkin & Goodenough, 1981), rod and frame (Corbett & Enns, 2006), and plum-line (Robert & Harel, 1996) tasks (Quaiser-Pohl et al. 2004; (Myer & Hensley, 1984; Signorella & Jamison, 1978); (Liben, 1978; Sholl, 1989); and (Barsky & Lachman, 1986; Corbett & Enns, 2006), respectively). Researchers suggested that performance on all these cognitive tasks correlates with an individual's field dependence, i.e. the extent to which an individual can dis-embed one part of a figure from its immediate surroundings.

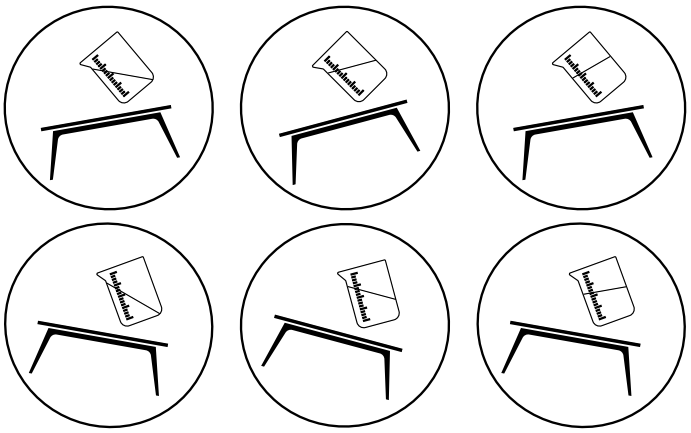


Figure 1.4: An illustration of photos with correct and anomalous water orientations used by McAfee & Proffitt (1991). The participants were asked to judge whether the pictures were anomalous or natural. The anomalous water was tilted 10° in the same or opposite direction of the container tilt (right and left columns, respectively). The pictures were cut in a circular shape and tilted +20° (top) and −20° (bottom), so the participants could only use the horizontal cues in the photos. This figure is created based on (McAfee & Proffitt, 1991).

Moreover, Barhorst-Cates et al. (2020) found children who drew their neighbourhood from a bird eye's view performed better on the WLT than those who drew it from their perspective. However, Quaiser-Pohl et al. (2004) reported that the correlation between the neighbourhood drawing style and performance on the WLT was not strong.

Error Types

Some researchers reported that adult errors differed from child errors in that adults did not typically draw lines parallel to the bottom of the container (Thomas & Jamison, 1975; Vasta & Liben, 1996; Vasta et al., 1993); however, other researchers observed such errors in college students (Howard, 1978). The direction and magnitude of the erroneous lines' tilts

are often left out of analyses. Vasta & Liben (1996) suggests that participants who reported they imagined water in a moving state drew lines that were tilted towards the rim of the container (reported from Vasta (1994)).

1.2.5 The Water-Level Task Variants

The conventional form of the WLT uses two-dimensional outline drawings of the same container in various orientations. However, in later studies, researchers used modified versions of the original Water-Level Task. This section summarizes different adaptations of WLT used in previous research.

Two Dimensional Drawings

As in the original WLT, in many WLT variants, the participants are asked to draw the water level in a tilted container. Researchers have used different types of containers (e.g. an hourglass-shaped container or a watering can (Robert & Harel, 1996). Additionally, some variants include visual cues to emphasize the stillness of the container and the liquid (e.g. a tilted bottle is mounted on a parked bicycle, or a baby bottle is securely held inside a warmer (Robert & Morin, 1993). Some researchers have used realistic photos rather than 2D outline drawings (McAfee & Proffitt, 1991).

Adjusting Water Level in Three-Dimensional Apparatus

Thomas et al. (1973) used a three-dimensional apparatus with a bottle and an adjustable disc representing the artificial water line. The participants could also look at a similar bottle half full of water before deciding. They found that many adults did not conceptualize that the water remained horizontal even when they could look at an actual container half full of water.

Adjusting Picture Orientation in Three-Dimensional Apparatus

McAfee & Proffitt (1991) created images of containers filled with liquid with no background cue and asked participants to adjust the images, so they looked upright. This variant drew the participants' attention to the environment as they could not change the water orientation relative to the container.

Identifying Anomalous Outline Drawings, Pictures, and Videos

Researchers created outline drawings (Vasta et al., 1996), images (Howard, 1978; McAfee & Proffitt, 1991), or videos (Howard, 1978) in which the water-level was not horizontal. They asked the participants to judge whether a stimulus looked natural or which of two stimuli looked correct.

Other Horizontal or Vertical Tasks

The Water-Level Task can be considered a subset of Horizontal/Vertical Tasks, all of which require using horizontal and vertical frames of reference and accounting for gravity. The Plumb-Line Task is an example of a vertical task in which participants are asked to draw a thread hanging from a standpoint with a plumb attached to it (e.g. A plant hanging from a rope attached to the ceiling of a house). In another example, participants must determine the flow of sand in a tilted hourglass. Researchers have tested variations of these tasks with the WLT and observed that the performance of individuals in these sets of tasks are highly correlated (Robert & Harel, 1996; Robert & Morin, 1993).

1.3 Intuitive Physics

Humans constantly interact with their environment and encounter physical principles underlying the world (e.g. gravity and movement principles). They must make judgements about

physical properties of the objects (e.g. weight, flexibility, speed), previous and next states of physical scene, and how their actions affect the world (Alex & Fischer, 2020). People's intuitive sense of physics does not always reflect their knowledge of physical principles (Proffitt & Gilden, 1989); however, it can affect their approach to solving formal physics problems (Simon & Simon, 1978).

Studies have shown that, from an early age, humans form an understanding of physical concepts. A human child as young as three months old understands that the world consists of bounded objects (Kestenbaum et al., 1987) and that the objects' motion is continuous in time and space (Spelke et al., 1995). By the age of one, most humans understand causal events (Oakes, 1994), have a rough understanding of the center of the mass and can predict the stability of an object put on the edge of another object (Baillargeon, 1996). Moreover, five-month-olds can differentiate liquid and solid behaviours; they expect that a liquid would pass through a grid and a solid object would not (Hespos et al., 2016).

Although from an early age, humans understand the physical world that surrounds them, their judgement of physical events is prone to error (e.g. McCloskey 1983; Proffitt & Gilden 1989; Sanborn et al. 2013). Intuitive physics researchers have used many physical scenarios to identify and explain human's intuitive understanding of physics and identify the perceptual, cognitive and analytical mental processes that are engaged in analyzing the physical world (J. R. Kubricht et al., 2017). Most research on Intuitive Physics used scenarios involving Newtonian mechanics. However, some researchers expanded their analysis to humans' understanding of liquid dynamics (Bates et al., 2015, 2019; Kawabe et al., 2015; Smith et al., 2013).

In the rest of this chapter, we will outline tasks used in studies of intuitive physics, summarize common findings in different tasks, and discuss models and explanations proposed by researchers about humans' intrinsic understanding of physics.

1.3.1 Tasks

Researchers have used various physical events in diverse settings to investigate human's intuitive understanding of physics. Although many observations were task-independent, we find it helpful to outline the tasks as it can give the reader ideas for future experiment designs. This section provides examples of some commonly used tasks and task conditions in Intuitive Physics research.

A large subgroup of studied tasks involves movement principles in particle dynamics (i.e. objects can be represented by their center of mass). These tasks include predicting the trajectory of moving objects in different settings. Proffitt & Gilden (1989) argued that people's understanding of particle dynamics is more precise than their understanding of extended body dynamics, where the distribution of the mass of objects affects their movements. They stated that for human intuitive physics perception to be accurate, only one salient feature (e.g. mass, velocity, volume) must affect the physical event, and that feature must be recognized correctly and not confused with irrelevant information.

In this section, we provide four examples of tasks that involve one relevant parameter, then discuss four examples that require more than one parameter to be considered simultaneously. Finally, we summarize three studies that involve the human understanding of liquid behaviour.

Object Falling from a Moving Carrier

Studies have shown that many people make mistakes when drawing the trajectory of an object falling from a moving carrier. Different carriers were used in a pen-and-paper version of the task, including a walking person dropping a pen, an object dropped from an airplane or abstract depictions of carriers and objects. Many people draw a straight line, while the correct answer is a parabolic trajectory that accounts for the objects' initial horizontal

momentum (McCloskey et al., 1983). McCloskey et al. (1983) observed that among 99 undergraduate participants, 49% drew a straight-down trajectory for a pencil dropping from a walking man's hand.

Object Falling from an Edge with Initial Momentum

Researchers observed that people make much fewer errors when a moving object falls from an edge with constant speed, although the physics is similar to the object being dropped from a moving carrier. In McCloskey et al. (1983)'s experiments, 94% of 34 undergraduate students drew a forward trajectory in an abstract representation of a ball falling down an edge.

Continued Curved Path in Absence of an External Force

Newtonian mechanics indicate that an object will move straight when released from a curved path in the absence of an external force. However, many predict it will continue moving on a curved trajectory. Examples include a ball rolling out of a horizontal curved track or an athlete throwing a disc. McCloskey et al. (1980) reported that nearly half of 50 undergraduate students drew a curved path in similar pen and paper scenarios. There is evidence that this misconception might affect one's actions too. McCloskey (1983) mentions that some people threw a ball on a table in a curved path, believing it would continue on a curved trajectory.

Pendulum Bob Movement when the String Breaks

Predicting the trajectory of a pendulum bob is similar to a projectile released from a curved path in that one must take into account the instantaneous velocity of the bob at the moment the string breaks. The bob's velocity is always perpendicular to the string, and its magnitude is zero at the top (apex) and maximum at the bottom (nadir). It is also similar to a falling

object because gravity affects its trajectory. As in the previous tasks, research has shown that people do not always draw a correct trajectory for the pendulum bob in a pen-and-paper version. Caramazza et al. (1981) asked 44 undergraduate students to determine the bob's trajectory if the string brakes in apex, nadir and between the two (Figure 1.5). Only 24% drew the correct trajectory for all states.

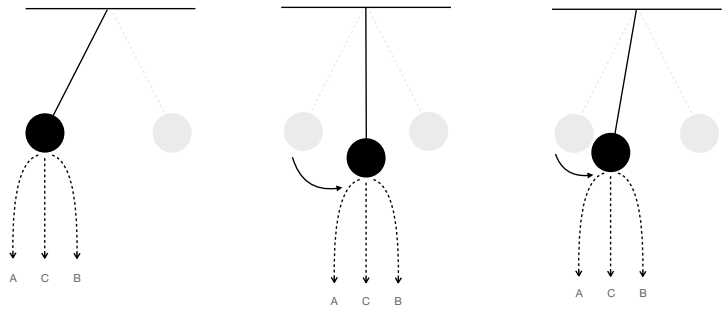


Figure 1.5: An illustration of trajectory prediction for pendulum bob. The participants are informed that the pendulum is swinging back and forth and are asked to choose between A, B, or C trajectory if the pendulum string is cut at that exact moment. The pendulum bob is at the apex, nadir and between the two in the left, middle, and right pictures, respectively. This figure is created based on Caramazza et al. (1981).

Collision

The collision task was initially created by Michotte (2017) to investigate humans' perception of causality. In the initial form, the participants viewed two-dimensional simulations of a moving object hitting another and judged if the collision caused the stationary object to move or if the object moved on its own (Sanborn et al., 2013). A typical adaptation of the task has been frequently used for Intuitive Physics studies, in which the participants see a simulation of two objects colliding and are asked to determine which object is heavier. Based on Newtonian mechanics, with or without loss of kinetic energy, the heavier object is the one

with the smaller change in velocity before and after the collision (Cohen, 2006; Cohen & Ross, 2009; Sanborn et al., 2013). However, researchers have observed that people's judgments are sensitive to irrelevant parameters. For example, a smaller restitution coefficient (i.e. a coefficient between 0 and 1 that determines the proportion of the kinetic energy observed after the collision) resulted in less accuracy (Todd & Warren Jr, 1982).

Extended Body Motion

The mass distribution of objects affects their angular velocity, so analyzing rotation requires extended body dynamics. For example, when spinning skaters open their arms, they spin slower. Proffitt et al. (1990) tested 12 undergraduate students in a simulated satellite game in which they could open and close panels of a spinning satellite affecting its angular velocity, and they were asked to judge if the simulation was accurate. The participants found the simulations in which opening the satellite's panels increased, decreased, or did not change the satellite's angular velocity equally natural.

Volume Displacement

"Suppose a stone is placed in a boat floating on water in a tank and the water level is marked. Will the water level change if the stone is taken from the boat and sunk into the water?" This is the volume displacement task in its complex form. It is known that even some renowned physicists mistakenly predict that the water level will not change (Kaiser et al., 1992; Proffitt & Gilden, 1989). However, the correct answer is the water level will decrease, because when the stone is floating, it displaces a volume of water equivalent to its mass, and when sunk, the displaced water equals to its volume (and the stone is denser than water). In simple versions of the problem, the water level is compared for two floating or two submerged objects differing in mass or volume or both. The relevant variable is mass in the floating tasks and volume in the submerged tasks.

In one of their experiments on 48 college students, Kaiser et al. (1992) asked participants to determine water displacement in complex and simple problems and showed them 2D diagrams of the task. They observed that in simple versions where only the relevant information was represented to the participants, the answers were correct 78% of the time. However, the number dropped to 48% when the irrelevant variable was also included. Only 21% of participants answered the complex problem correctly, 39% percent said the water level would remain at the same height, and 40% said it would increase.

Unstable Towers

The unstable towers task asks people to judge whether or not a tower of blocks is stable or predict in which direction it will fall (Alex & Fischer, 2020; Battaglia et al., 2013; Fischer et al., 2016). Research showed that people are relatively accurate in their predictions. For example, Alex & Fischer (2020) observed people predicted the direction that towers would fall correctly 75% of the time in 3D synthetic pictures. This task differs with previous tasks first in that it requires rough estimation and not exact reasoning, and second in that people are asked to determine the state of the system in a single point in future rather than elaborate explanation of what happens in each moment. Alex & Fischer (2020) found high correlation between performance on unstable towers tasks and spatial cognition tasks (mental rotation (Shepard & Metzler, 1971) and paper folding (Ekstrom & Harman, 1976).

Liquid Flow

Bates et al. (2015) studied people's intuitive physical reasoning about liquids. They presented participants with synthetically generated images of a mass of water above some obstacles and asked them to predict, if released to fall under gravity, what percent of the water would end up in the left and right half of the image. They compared human predictions with various simulation and prediction methods. They used Smooth Particle Hydrodynamic

(SPH) liquid simulation method (Monaghan, 2005) as the ground truth. They observed that human performance was best described by SPH simulation if an Additive white Gaussian noise was added to the initial position of each particle.

Whether or not this study analyzed the general human understanding of liquid dynamics is arguable because the participants went through a training phase in which they viewed how the simulated liquid would behave in a similar situation. Still, it is interesting that the participants' estimations could be modelled by adding noise to the simulation model.

Water Pouring

The water pouring task is similar to the Water-Level Task as it requires imagining liquid in tilted containers. In the water pouring task the participants are asked to make judgements about how much an upright container containing water (or other liquids) has to rotate for the water to start to pour out. Schwartz & Black (1999) showed participants actual empty containers with water line marked on them. They asked participants to rotate or imagine rotating the containers with closed eyes just enough for the water to pour out. Participants correctly tilted narrower containers more than wider containers (84% and 100% success for tasks in which the difference in the required tilts were 10 and 20 degrees, respectively), and their estimations were close to the actual required tilt. However, the participants were not accurate when asked to explicitly state which container they tilted more, and their explanations were not consistent with their actions. Less than 15% of participants correctly explained that the narrow container must be tilted more.

Also, Schwartz & Black (1999) observed that participants were not accurate in a pen-and-paper version of the task. In the pen-and-paper version, people saw two vertical containers, one of which had a water line, and were asked to draw the water line in the other so the water would pour out at the same tilt amount for both containers. The researchers also asked participants to explain their strategy for drawing the line. The results showed that people

incorrectly drew higher lines for wider glasses and vice versa. Moreover, it was observed that participants' strategies were not consistent across different trials and did not correlate with their performance in the tasks.

J. Kubricht et al. (2016) designed a liquid-simulation-based study inspired by Schwartz (1999)'s observation that when tilting a container with imagined liquid inside so the liquid will be about to pour out, participants' actions were affected by the imagined liquid's viscosity. In the experiment, participants viewed computer-generated animations of two liquids with different viscosity pouring into two containers. Then the participants saw computer-generated images of two similar containers filled with the two liquids to different or similar heights and were asked to determine "which container will need to be tilted with a larger angle before the fluid inside begins to pour out". The results showed that participants' answers were sensitive to the liquids' viscosity as well as the liquids' heights. Especially in cases where the height difference was small, the participants tended to think the viscous liquid needed to be tilted more to pour out. This observation suggests that people did not use a simple heuristic (i.e. "The lower liquid must be tilted more") in solving all the trials. Similar to Bates et al. (2015), J. Kubricht et al. (2016) used a particle-based liquid simulation method (Fluid Implicit Particle/Affine Particle in Cell (FIP/APIC) (Jiang, 2015)), added noise to its calculations, and showed its performance could approximate the overall participants' performance.

1.3.2 Common Observations

Some findings and patterns have been consistently repeated across different tasks and studies. As several examples in section 1.3.1 show, humans usually perform better in task settings that require action and implicit reasoning compared to explicit descriptions or abstract illustrations settings. For example, participants' actions were approximately consistent with

physics principles when asked to throw a ball (Krist et al., 1993) or tilt a container so the liquid would pour out (Schwartz & Black, 1999). However, they were not accurate in explaining their actions. Moreover, Smith et al. (2013) found that subjects performed more accurately on cutting a pendulum string to hit a target in a simulation (acting) than drawing the bob's trajectory (abstract illustration). Additionally, people make more accurate judgements if the task context is familiar. Most people indicated that water going out of a curved hose continues straight; however, when asked about a ball going out of the same hose, many predicted a curved trajectory (Kaiser et al., 1986). It has also been reported that human predictions are sensitive to minor changes in problem presentation. For example, in the WLT, people's success depends on the container's shape and tilt degree (Pascual-Leone & Morra, 1991).

Animation, Video, and Real-Life Illustrations

Kaiser et al. (1992) proposed that dynamic presentation improves one's predictions and judgments of physical events if two conditions are met. First, the physical event is simple, meaning only one "dimension of information" affects how the events unfold, and second, the animation draws one's attention to that information. In one of their experiments, Kaiser et al. (1992) created natural and anomalous videos of the complex water displacement task (section 1.3.1) using a bolt floating on a boat in a water tank. The water tank design allowed altering the amount of water in the tank without it being noticeable in the videos. They created videos in which the bolt was removed from the boat and submerged in the water. In some videos they altered the natural water level when the bolt was put in the water. They asked 12 participants to view pairs of video tapes and judge which was more natural. Eleven out of 12 participants correctly chose the videos in which the water level decreased when the bolt was sunk. Participants were significantly more accurate when watching video than when making judgments in an abstract settings. Kaiser et al. (1992) also observed

that participants could identify anomalous pendulum bob trajectories in simulations. On the other hand, animating a satellite's movements did not help participants identify its anomalous rotation when its mass distribution changed (Proffitt et al., 1990).

Individual Inconsistency

As discussed before, an individuals' explanations and abstract illustrations are not always consistent with their actions. Schwartz & Black (1999) observed that participants explanation of their strategies were not consistent across different trials. Also, participant errors were not always consistent in different trials.

Knowledge and Training

It has been observed that knowledge of physics principles correlates with one's performance in the tasks. For example, (Liben & Golbeck, 1984) observed that mentioning the horizontality principle to participants significantly improved their performance in WLT. Also, many individuals learn from explicit (e.g. feedback (Cohen, 2006)) or implicit (e.g. seeing tasks from easy to hard (Vasta et al., 1996)) training and change their strategy. However, their modified tactics might not align with physics principles; in other words, individuals might learn and use simplistic heuristics from training examples that only capture some aspects of the physical events (Cohen, 2006).

Knowledge of a physics principle does not always affect one's intuitive physics perception. For example, as discussed in section 1.2.2, some individuals failed the WLT despite identifying the horizontality principle (Myer & Hensley, 1984). Also, Proffitt et al. (1990) observed that high-school teachers and physics PhD students did not have a correct physics understanding of wheel movements. Moreover, participants do not always transform knowledge and observation from familiar contexts to abstract or less familiar contexts, meaning observation does not always result in explicit knowledge. For example, viewing images of

tilted containers containing liquid did not affect participants subsequent performance on WLT (Thomas et al., 1973).

1.3.3 Intuitive Physics Models

Fischer et al. (2016) observed that distinct cortical regions in the human brain, "bi-lateral frontal regions (dorsal premotor cortex/supplementary motor area), bilateral parietal regions (somatosensory association cortex/superior parietal lobule), and the left supramarginal gyrus", were active specifically when making judgements requiring physical analysis –using variants of the stable towers (section 1.3.1) and collision (section 1.3.1) tasks– or even when passively viewing physical events. The active areas were consistent among subjects and tasks and overlapped with areas known for motor action planning, tool use, problem-solving, and cognitive reasoning (Fischer et al., 2016). Similar regions have been observed to be active while viewing scenes that include visual information about objects' weight (Loh et al., 2010). These findings suggest that the human brain has specific resources for processing physics-related information. However, how these resources operate is the subject of debate. More specifically, it is unknown whether humans have intrinsic knowledge of physics and if a unified process is used in making decisions in different scenarios and contexts.

Some researchers believe that naive and inaccurate heuristics can explain misconceptions about physical events. Piecemeal Heuristics models describe the humans' physics understanding as a collection of task-specific heuristics learnt through experience. For instance, in collision tasks, simple heuristics such as making a decision based on objects' final velocities or whether or not one object ricochets could describe human choices and individual differences (Cohen, 2006). Also, in the WLT, some individuals use the inaccurate heuristic that the water remains parallel to the bottom of the container.

Perceptual biases might cause errors in the learnt heuristics. For example, McCloskey

et al. (1983) hypothesized that the straight-down belief for the moving carrier task (section 1.3.1) originates in a visual illusion. When one drops an object while moving, the dropped object's motion relative to the moving person is straight down. Their studies also indicated that even for a stationary observer, the perceived trajectory of a falling object is dramatically affected by the presence of a second moving object (e.g. the carrier). Another example of such biases is the effect of tilt illusion on perceiving the water level in a tilted container discussed in section 1.2.4.

It has been argued that in many tasks, people's errors and strategies are not consistent across different trials and their explicit explanations do not match their performance (Schwartz & Black, 1999), thus the error cannot be attributed to an established but inaccurate understanding of the underlying physical principle. Moreover, in some cases, alternative models have been shown to describe overall human behaviour more accurately than the Piecemeal Heuristic models (Bates et al., 2015; Sanborn et al., 2013).

Probabilistic and Simulation-based models have been proposed as alternatives to the Piecemeal Heuristic models. These models assume that the human brain can simulate the physical world rather accurately, and our actions are based on simulating and predicting the world's state in the future. In these models, human errors originate from the noisy observations of the state of the world. As an example, the Noisy Newton model assumes the human brain calculates the future using precise physics formulae; however, our decision is based on a predicted future state with maximum probability given our noisy observations of the world (Sanborn et al., 2009, 2013). Similar to Noisy Newton, the Simulation-Based model proposes that the human brain performs a probabilistic simulation of physical phenomena to predict the outcome of various events (Battaglia et al., 2013; Hegarty & Sims, 1994; J. Kubricht et al., 2016). However, it assumes that rather than using precise physics formulae, our brain uses simplified physics principles to simulate the state of the world in a step-wise manner (Hegarty, 2004), in a similar manner to that used by physics engines used in video

or gaming simulations (Bates et al., 2015; Battaglia et al., 2013; J. Kubricht et al., 2016).

Despite the similarities between computer-generated probabilistic physics simulations and human predictions, reasoning and prediction are sometimes more efficient using heuristics (Davis & Marcus, 2016). For example, humans effortlessly infer that water remains inside a closed container. They do not have to simulate or imagine the state of water particles. Thus it is possible that reasoning about physical events relies on both learnt heuristics and mental simulation (Hegarty, 2004).

The nature of the Intuitive Physics tasks examined in evaluating the two main models has been different. The Piecemeal Heuristic research has primarily used tasks involving single variables (e.g. trajectory prediction tasks). In contrast, research on Simulation-Based models has focused on more complex tasks (e.g. liquid flow prediction and unstable towers). The human performance in the collision task has been modelled with both approaches.

Summary

Neither of the models discussed above can fully explain all of the common observations in humans' performance in Intuitive Physics tasks. The heuristic model can describe errors commonly made by people, especially in abstract tasks, and it can explain individual differences. However, it cannot justify individuals' inconsistent performance in different trials and humans' accurate predictions in complex scenarios. The Simulation-Based models successfully make predictions similar to the average responses of humans, but they do not account for individual differences. Also, they can not explain why small changes in task presentation highly impact one's performance (Kaiser et al., 1986). Also, it is evident that people change their answers to physics problems when given explicit or implicit feedback or training (Ranney & Thagard, 1988; Vasta et al., 1996). However, the learning process and how extrinsic knowledge affects one's intuitive physics understanding are still unclear to researchers.

As an additional note, the author believes that, in some cases, the inconsistency of human

predictions and Newtonian physics can be due to leaving out factors such as friction and air resistance when determining the correct answer for the tasks. However, in real life, they affect the outcome of similar events. For example, a spinning projectile's trajectory is curved when it moves through the air (known as the Magnus effect), or air resistance can cause a dropped object's projectile to be close to a straight line (i.e. dropping a light tissue while walking).

1.4 Statement of Purpose

Although the Water-Level Task has been extensively studied as a spatial cognition task, there is more to be discovered about it as an intuitive physics experiment. It has been observed that from an early age, humans have a general understanding of how a liquid behaves (Hespos et al., 2016), and adults can infer complex concepts, such as viscosity, from motion videos (Kawabe et al., 2015). It is intriguing that despite proficiency in determining complex behaviours, the simple heuristic that the water remains horizontal is only sometimes known to individuals.

From an Intuitive Physics perspective, the Water-Level Task is simple because it only involves one parameter of interest, gravity. With the same reasoning as Proffitt & Gilden (1989), the common error observed in the WLT can be attributed to irrelevant parameters, such as the tilt of the container, drawing attention away from the relevant parameter. Proffitt & Gilden (1989) suggest animation might improve people's performance in simple physics problems as it could draw their attention back to the relevant parameter. However, studies have reported that using dynamic images did not improve performance on WLT (Howard, 1978; McAfee & Proffitt, 1991).

This project will study the Water-Level-Task in an abstract (pen-and-paper) form and in a more realistic setting (augmented reality). Although the AR environment is less abstract

than 2D depictions, it involves more irrelevant parameters that might draw an individual's attention away from the relative parameter. Despite this, the AR environment might improve participants' performance because it has the everyday horizontal and vertical cues and it also allows interacting with the simulations. Schwartz (1999) observed that participants were able to predict how much a container must tilt so an imaginary liquid would pour out. The AR-WLT interaction allows participants to tilt the containers and use such visual cues to identify anomalous simulations.

Additionally, we will analyze the performance of inaccurate participants across the pen-and-paper and AR experiments to discover whether or not a similar strategy is involved in their physics apprehension and abstract decision making process.

Finally, a few intuitive physics studies have analyzed humans' interactions with anomalous simulations. Our AR-WLT platform allows such experiments, which allows further investigation of how anomalous physics affects human performance. This is helpful in understanding human mental processes involved in everyday interactions with the physical world.

Chapter 2

Augmented Reality Liquid in Container

2.1 Introduction

This chapter discusses the implementation details of the Augmented Reality (AR) simulation of water in a container. The AR liquid simulation was implemented in Unity (Unity Technologies, 2022) and deployed to a HoloLens2 (Microsoft, 2022) device. We will discuss the two main components, the liquid-in-container effect and human interaction with the container.

2.2 Liquid-in-Container Simulation

The state-of-the-art approaches for simulating liquids are particle-based methods which require a high number of simulated particles to look realistic, so a robust graphics processing unit is required for rendering in real-time applications. We tested two implementations of particle-based methods in Unity and on a computer with an Nvidia GeForce GTX 1050

GPU, Intel® Core™ i7-9700 CPU, and a 16 gigabytes memory. Even on the computer, both packages were slow, so we implemented a simple liquid effect in Unity, which is less realistic but is light in computation and suitable to use in the HoloLens2 device. This section briefly summarizes state-of-the-art liquid simulation packages we tested. Then we discuss our liquid effect in detail.

2.2.1 Existing Liquid Physics Simulation Methods

We tested two liquid simulation assets in Unity: Obi Fluids (Virtual Method Studio, 2022) and Zibra Liquids (Zibra AI, 2022). We faced two main problems with these approaches; first, even with the computer's processing resources, the number of particles had to be set to a small value to get to 20-30 frames per second in the simulation, resulting in a bumpy, jelly-looking liquid surface. Second, when the particles were inside a movable container, if the container moved faster than a threshold and the particles were small, the liquid leaked out of the container's surface; This is known as the "Tunneling effect"[1].

2.2.2 The Liquid-in-Container Effect

We implemented a liquid-in-container effect in Unity using a method used in real-time games, such as the Virtual Reality game "Half-Life: Alyx" (Valve Corportation, 2020). In this method, the liquid effect is achieved by only rendering the container pixels that are below an imaginary plane, which is perceived as the liquid surface[2]. The liquid surface oscillates proportionally to the container's translation and rotation speed, which results in a liquid feeling. The remainder of this section summarizes the details of the implementation, its shortcomings and future directions for improving the liquid effect.

[1]This is mentioned in Obi Fluid's official forum `http://obi.virtualmethodstudio.com/forum/archive/index.php?thread-1411.html`

[2]I used this blog post as a reference: `https://80.lv/articles/simulating-liquids-in-a-bottle-with-a-shader`

Liquid In Container Implementation Details

The liquid effect simulation consists of two components, a vertex-fragment shader and a C# script.

The vertex-fragment shader gets a 3D vector, $\overrightarrow{\mathbf{n}}$, as input. The liquid surface plane is described with $\overrightarrow{\mathbf{n}}$ as its normal and a point on the plane in world coordinates. We use the middle point of the container as the point on the plane, so the liquid plane always goes through the middle point of the container, which results in the container being half full all the time. The middle point of the container in world coordinates, $\mathbf{o}_w$, can be calculated by converting the origin in object coordinates, $\mathbf{o}_o = (0, 0, 0)$, to the world coordinates: $\mathbf{o}_w = ObjectToWorldMatrix \times \mathbf{o}_o$. The shader uses the plane equation (equation 2.1) to determine the location of each container surface pixel $\mathbf{p}$ relative to the liquid surface. It renders the pixels that are on or below the surface.

$$\text{liquid surface plane equation} : \overrightarrow{\mathbf{n}} \cdot \mathbf{x} - \mathbf{o}_w = 0 \tag{2.1}$$

$$\text{Shader output for pixel at location } \mathbf{p} = \begin{cases} \text{pixel's color} & \text{if } \overrightarrow{\mathbf{n}} \cdot \mathbf{p} - \mathbf{o}_w <= 0 \\ \text{do not render} & \text{otherwise} \end{cases} \tag{2.2}$$

Finally, we insert the cylinder with the liquid effect shader inside a slightly bigger cylinder with a transparent shader to mimic a transparent container containing the liquid. The final shader in a cylindrical container is shown in figure 2.1. In this figure, what is perceived as the liquid surface color is the pixels on the rear side of the cylinder rendered with a darker color. Figure 2.2 shows the shader used on a cylindrical container with different plane normals.

The C# script is responsible for controlling the liquid normal to mimic the fluid surface movements when the container moves or rotates. When the container is still, the liquid plane normal is always in its stationary state, which is the opposite direction of the gravity:

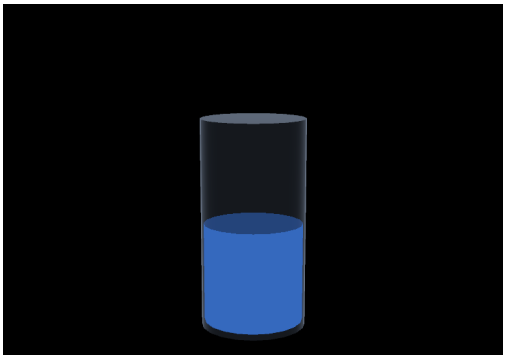


Figure 2.1: The final liquid in container shader effect.

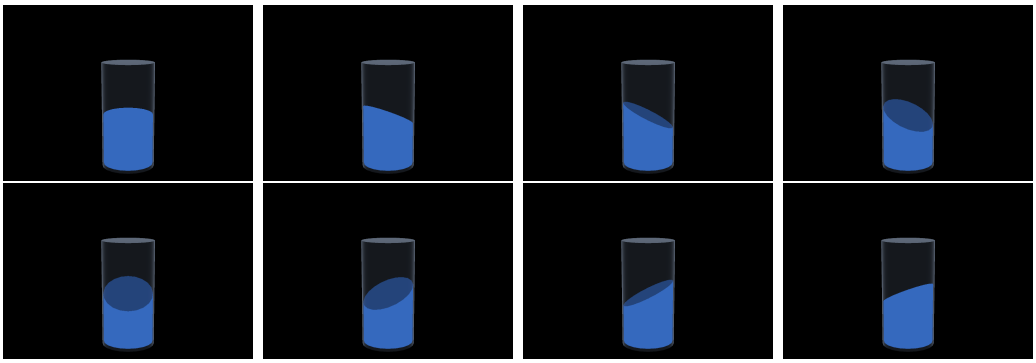


Figure 2.2: Liquid shader used on a cylindrical container with plane normals $\overrightarrow{\mathbf{n}} = (0, 0.90, 0.45)$, $\overrightarrow{\mathbf{n}} = (0.31, 0.9, 0.31)$, $\overrightarrow{\mathbf{n}} = (0.45, 0.9, 0.0)$, $\overrightarrow{\mathbf{n}} = (0.31, 0.9, -0.31)$, $\overrightarrow{\mathbf{n}} = (0.0, 0.9, -0.45)$, $\overrightarrow{\mathbf{n}} = (-0.31, 0.9, -0.31)$, $\overrightarrow{\mathbf{n}} = (-0.45, 0.9, 0.0)$, $\overrightarrow{\mathbf{n}} = (-0.31, 0.9, 0.31)$, from top left to bottom right.

$$\overrightarrow{\mathbf{n}} = -g = (0, 1, 0). \tag{2.3}$$

When the container moves or rotates, the liquid plane oscillates about its stationary state. The oscillations are calculated in $\mathbf{x}$ and $\mathbf{z}$ directions separately. At each frame t, the liquid normal is calculated as $\overrightarrow{\mathbf{n}} = \overrightarrow{\mathbf{n}}_t / |\overrightarrow{\mathbf{n}}_t|$, where $\overrightarrow{\mathbf{n}}_t = (x_t, 1, z_t)$, and x_t and z_t are sinusoidal waves calculated using equations 2.4 and 2.5.

$$x_t = \alpha x_{t-1} + A_x \cos(2\pi f t - \phi) \tag{2.4}$$

$$z_t = \alpha z_{t-1} + A_z \cos(2\pi f t - \phi) \tag{2.5}$$

In the above equations f is fixed and set to 3, so the water plane oscillation rate is 3 cycles per second. A_x and A_z are calculated based on the container's acceleration $\overrightarrow{\mathbf{a}_t} = (a_{tx}, a_{ty}, a_{tz})$. and angular velocity $\overrightarrow{\boldsymbol{\omega}_t} = (\omega_{tx}, \omega_{ty}, \omega_{tz})$ at frame t as follows:

$$A_x = \beta a_{tx} - \gamma \omega_{tz} \tag{2.6}$$

$$A_z = \beta a_{tz} - \gamma \omega_{tx} \tag{2.7}$$

The coefficient $0 \leq \alpha < 1$ is a damping factor and is set based on delta time (Δt) at each frame update ($\alpha = max(1 - 2 * \Delta t, 0)$). At each time frame t, new sinusoidal waves are added to the liquid plane's oscillation in x and z directions, whose amplitude is proportional to the container's angular velocity and acceleration in corresponding directions. Hyper-parameters β and γ were chosen heuristically. The phase of the newly added sinusoidal waves, ϕ, is set to $2\pi f t$, so the new wave peaks at time t.

As equations 2.4 and 2.5 indicate, at each time t, x_{t-1} and z_{t-1} are also sinusoidal waves with the same angular velocity $2\pi f$ and different amplitude and phase. Thus, x_t and z_t are sinusoidal waves with angular velocity $2\pi f$, and only the amplitude and phase of each wave update at each time frame. The new waves x_t and z_t are calculated using the equation for adding two sinusoidal waves with the same frequency:

$$A_1 \cos(2\pi f t + \phi_1) + A_2 \cos(2\pi f t + \phi_2) = A_3 \cos(2\pi f t + \phi_3), \tag{2.8}$$

where

$$A_3 = \sqrt{[A_1 \cos(\phi_1) + A_2 \cos(\phi_2)]^2 + [A_1 \sin(\phi_1) + A_2 \sin(\phi_2)]^2}$$

and

$$\phi_3 = \arctan\left(\frac{A_1 \sin(\phi_1) + A_2 \sin(\phi_2)}{A_1 \cos(\phi_1) + A_2 \cos(\phi_2)}\right).$$

The Liquid-in-Container Effect Limitation and Constraints

The method described in section 2.2.2 creates a simple, fast liquid effect. Because of its simplicity, it has some limitations. The main limitation is that the method is a simple liquid "effect" and does not accurately simulate liquid surface physics. Aside from disregarding splash and sloshes on the liquid surface, our method can be inconsistent with liquid physics in cases such as the container moving with a constant acceleration $\overrightarrow{\mathbf{a}}$ in the x direction. In this case, the liquid surface must remain at a fixed tilted position ; however, in our simulation, the liquid surface keeps oscillating. Moreover, in our method, the container's angular velocity has a direct effect on the liquid surface's oscillations. In real physics, a container's rotation affects the liquid surface movement in various ways based on the rotation axis, container shape, fill amount, and gravity. For example, if the container rotates about an axis parallel to the gravity (in our case, the upright container rotates about its y-axis), the liquid surface curves towards the edges of the container.

Another limitation arises because the method does not render any vertices on the liquid surface, and the pixels of the rear side of the container are perceived as the liquid surface. As a result, we cannot add texture to the liquid surface to achieve a reflective, more realistic-looking surface. To solve this, some programmers project the vertices above the liquid plane on the plane (instead of not rendering them). However, it is not straightforward which direction to use for the projection and how to keep the projected vertices inside the container. Thus, for the projection to work, the container must be completely symmetrical (a sphere), or its rotation must be limited.

In addition to the limitations discussed above, our implementation only allows the container to be precisely half full of water, and it only works for a container that is symmetrical about its origin. As mentioned in section 2.2.2, we use the origin in container space as a point on the liquid plane. Thus, if a container is symmetrical about its origin, the liquid

plane always divides the volume of the container in half regardless of the liquid plane normal, so the liquid volume is preserved. If the container is not symmetrical or the point on the liquid plane is not the symmetry point of the container, the method fails to preserve the volume of the liquid.

Finally, this liquid effect only works for use cases in which the container is a closed shape, and the liquid does not flow out of the container. Although, it is possible to create a liquid flowing out effect using Unity particles and lowering the liquid plane, however, it should be done with caution as the container will not be half full anymore, so the origin of the container coordinate will not work as a point on the liquid plane.

To summarize, our "liquid effect" implementation is a simple approach that makes the users believe that the container is half full of liquid. Although it does not accurately simulate the physics of liquids, with some considerations, it can be used to examine human's understanding of liquid orientation in a tilted container.

2.2.3 Interaction With the Container

One major component in the AR implementation of liquid in a container is how users interact with the container in the AR environment. We wanted the interaction to be natural as possible. Our first attempt was to use Vuforia's (PTC Inc., 2022) object tracking implementation, which can be integrated into HoloLens2 via their unity asset. With this approach, a label would be attached to a real object in the world. Vuforia's tracking would follow the label's location and orientation and render the liquid container on top of the label. This approach would allow a tangible interaction with the container in the AR environment, in which the container would be affected by real world gravity.

However, we faced a significant problem: the rendered container disappeared if the tracking system could not follow the label. On the HoloLens2 device, if the label was rotated or

moved to extreme points or the movement was fast, it took time for the system to find the label and render the container on top of it, which resulted in an inconsistent rendering of the container.

As a result of these problems, the approach we used for participants' interaction with the container was using HoloLens2's built-in hand-tracking system. It can be integrated with Unity through Microsoft's Mixed Reality Toolkit (MRTK) (Mircosoft, 2022). With hand-tracking, virtual objects can be grabbed, moved, and rotated. However, real-world physics, i.e. gravity, does not affect the objects, which results in the objects remaining in their position when they are released. One advantage of this is that the participants can leave the containers at any position and with any orientation they want to evaluate the liquid inside the object. However, the interaction would be less realistic because the participants are interacting with a virtual object that does not have a real-world physical presence. Figures 2.3 and 2.4 show a user interacting with a container using HoloLens2 from the user's and an observer's point of view respectively.

Figure 2.3: User's observation when grabbing (left), manipulating (middle) and releasing (right) the container.

Figure 2.4: A user grabbing, manipulating and releasing the container from left to right.

Chapter 3

Method

3.1 Introduction

This project's main goal was to examine the Water-Level Task (WLT) in a more realistic setting. In other words, we were interested to know whether individuals who do not draw a horizontal line would find a simulation of liquid that does not remain horizontal more realistic than one that remains horizontal. To investigate whether the error happens in a more realistic setting, we designed an Augmented Reality Water-Level Task (AR-WLT). We used our Liquid-in-Container implementation (discussed in chapter 2) and altered the water surface's horizontality by adjusting the liquid-plane's stationary orientation, i.e. gravity direction. Then, we asked the participants to choose between two simulations with different stationary orientations of the water surface plane. As in previous research on individual differences in WLT (summarized in 1.2.4), first, we conducted the conventional WLT (prescreening) to recruit two balanced sets of individuals who were and were not susceptible to the conventional WLT error. Then, we tested the recruited participants in the AR-WLT.

Additionally, we added a control task, also done in AR, in which the participants were asked to adjust a virtual surface to make it horizontal. Only subjects recruited to participate

in the AR-WLT did the control task, which was done after the AR-WLT was completed. The control task was added to evaluate participants' ability to determine horizontal surfaces in the AR environment.

All participants were York University students enrolled in the PSYC 1010 course in the Summer of 2022 and participated in the experiment for course credits. The credits for prescreening and the AR-WLT were granted independently. The participants had to finish the prescreening to get the automatically granted credits. However, for the AR-WLT, the participants could take a rest or terminate the experiment at any point, and it did not affect their granted credits. The details of prescreening, AR-WLT, and the control task are discussed in sections 3.2, 3.3, and 3.4, respectively.

3.2 Prescreening

Unlike most previous research, our conventional Water-Level Task was conducted online. We used the JSPsych JavaScript library (De Leeuw, 2015) for implementing the user interface of the digital WLT and deployed the app to the Pavlovia (Bridges et al., 2020) server[1]. The participants could participate in the research at any time and use any device with an internet connection and a browser, including a smartphone, tablet, laptop, or computer, to access the study. The only limit was that their device had to have a higher resolution than 200×200 pixels. Depending on their device, they could complete the tasks using a touch-screen, trackpad, or mouse. We adjusted the trial images' size based on the participant's screen size so the complete image, including the outside black frame, would fit inside the participant's screen.

The conventional WLT consisted of 16 puzzles of 4 different containers, each of them appearing at four different angles. Figure 3.1 shows four puzzle examples. The participants

[1]Website: `https://pavlovia.org/`

saw puzzles one-by-one in random order and had to solve each puzzle to see the next one. In all the pictures, the container was depicted inside a black box and was resting on a horizontal line (surface), which could be used as a horizontal reference.

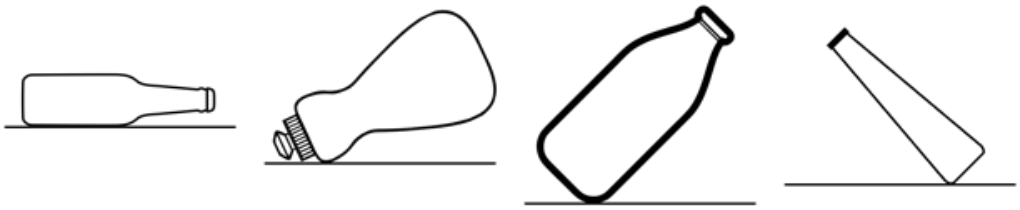


Figure 3.1: Examples of the conventional WLT puzzles (the "beer-bottle", "dish-soap-container", "milk-bottle", and "simple-bottle" from left to right).

As in the conventional task, in the first step of the online WLT, the participants were shown an upright container half full of water (figure 3.2).

Figure 3.2: The first step of the online WLT: a solved puzzle shown to participants.

In the second step, the participants were asked to draw a waterline in the same upright container they previously saw. This additional step was added to the conventional WLT as a tutorial to ensure the participants could use the online platform without hardship.

Finally, the puzzles were shown to participants randomly, and the participants had 20 seconds to solve each puzzle. Each puzzle had to be solved using one continuous line, and if the drawing input was not a continuous line, the app would show an error message to participants and repeat the same puzzle.

The online app saved each drawn line as a sequence of 2D points. To calculate the line's tilt from horizontal, we fitted a straight line to the 2D points using the Least Squares method and calculated the angle between the fitted line and horizontal to determine the tilt from horizontal in degrees. The whole process was done automatically using JavaScript and Python scripts. Figure 3.3 shows examples of participants' answers and the calculated fitted lines.

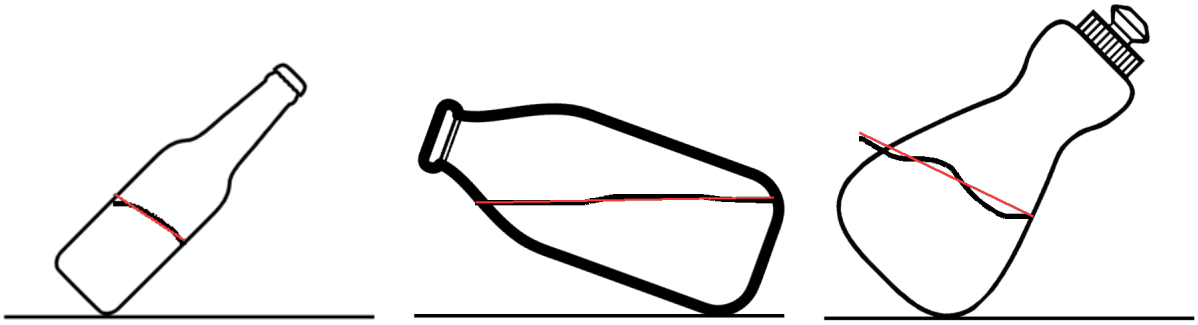


Figure 3.3: Three examples of participant's water-lines (black) and the lines fitted to their answers using Ordinary Least Squares (red).

For each puzzle, if the answer tilt was less than a threshold th, the puzzle was considered correct, otherwise it would be incorrect. We tried three different thresholds, 5°, 10°, and 15° and chose 15° for our analyses, which is less strict than the previously used 5° and 10° but provided better balance between high and low-scoring groups. The details of our analyses can be found in section 4.2. For each participant we calculated their total score as the number of correctly solved puzzles, which ranged between 0 and 16. We also saved participants' gender (chosen between the options: male, female, transgender, non-binary,

intersex, other, and prefer not to say), age (any number between 18 and 99), and the screen size of the device they used (as small (between 200×200 and 300×300 pixels), medium (between 300×300 and 600×600 pixels), or large (larger than 600×600)).

We recruited participants whose scores were on the two extremes of the score range to participate in the AR-WLT. Participants with scores less than or equal to 1 and more than or equal to 13 were invited to participate in part 2 as low-scoring and high-scoring participants, respectively. The total score thresholds for low-scoring and high-scoring participants were chosen based on our analysis results discussed in section 4.2.1.

3.3 The Augmented Reality Water-Level task

The AR-WLT was designed to investigate participants' perception of water orientation in a tilted container in an AR environment, where they could interact with the half-full containers. Specifically, we wanted to know if the water that does not remain horizontal in a tilted container appears less natural to the participants and investigate individual differences in perception of a realistic-looking water surface orientation.

AR-WLT was created using the Liquid-in-Container effect discussed in chapter 2, and the stationary orientation of the liquid-plane was altered by altering its normal. We asked participants to choose which of two simulations looked more realistic, in one or both of which the liquid-plane did not remain horizontal. In half of the trials (the covered trials), we added a white box cover around the container as soon as the participants grabbed it. Hence, the participants only saw the liquid surface when the container was released (more detail is provided in section 3.3.3). The covered trials were included to examine the participant's judgement of the water's stationary orientation without seeing the dynamic interaction.

In the remaining section, we first discuss how we altered the water orientation. Then, we will provide the details of the experimental procedure and trials, data, and evaluation of

participants' performance on the AR-WLT.

3.3.1 Water Orientation Adjustments

In the real world, water in a container remains horizontal (i.e. perpendicular to gravity) in a static state where no external force is applied to the container. So, in a stationary state, the water-plane normal is the negative direction of gravity, or $\overrightarrow{\boldsymbol{n}} = (0, 1, 0)$ in world coordinates. In the AR-WLT, we tilted the water-plane normal when the container was tilted to create anomalous orientations.

We altered the water-plane normal in a way to keep the water surface horizontal when the container was upright or at 90 degrees (i.e. lying flat). In other cases (when the container was tilted), the water would tilt proportional to the container's tilt. Moreover, we kept the water-plane normal above the horizontal plane, so the water did not look upside-down. Also, we ensured that the water-plane would rotate in the same or the opposite direction as the container's rotation. Formally, we rotated the water-plane about the same axis as the container's rotation axis to ensure the container's axis of symmetry (y-axis), the world coordinate's y-axis, and the water-plane normal would remain on one plane. As a result, the water surface always touched the container's rim in the correct spot. Figure 3.4 provides examples of water orientation adjustments. We will discuss the water-plane alterations in more detail.

We defined $0° \leq \theta_c \leq 90°$ as the container's tilt from upright (i.e. the smaller angle between the world's y-axis and the cylinder's axis of symmetry) and $\boldsymbol{\theta_c} = \theta_c \overrightarrow{\mathbf{e}}$ as the axis-angle representation of a rotation that rotates the container from an upright to the current tilt. We used an "anomaly coefficient", $-1 \leq a \leq 1$, to determine the magnitude of water orientation alterations. We rotated the water-plane $\theta_n = a \times \theta_c$ degrees about the same axis as the container's rotation up to a container tilt of 45° and gradually rotated the water

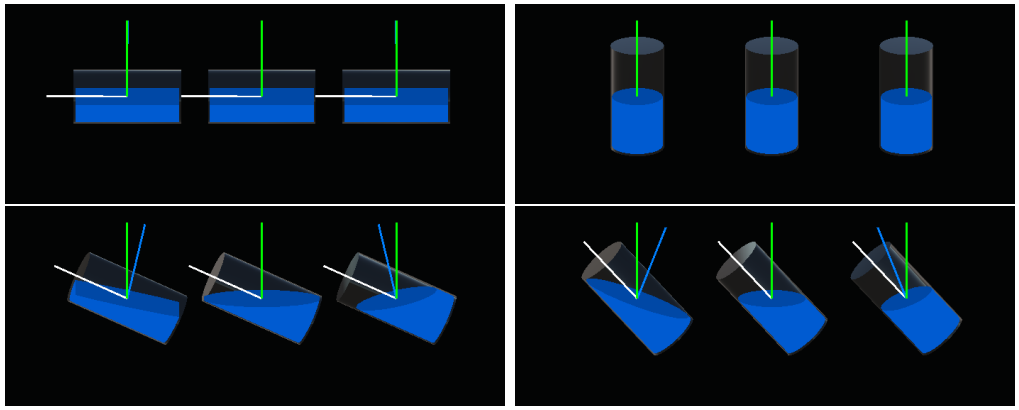


Figure 3.4: Examples of water plane alteration. The blue, white, and green lines show the water plane normal, the container's axis of symmetry, and the y axis in world coordinates, respectively. The water plane is tilted in the opposite and same direction as the container tilt in the left and right containers, respectively. The middle container shows an unaltered (horizontal) water plane. The top figures show that the modified water plane is horizontal when the container is upright and horizontal. The bottom figures show two examples of the water plane rotation when the container is tilted between 0 and 90 degrees.

surface back to the upright position when the container tilted from 45° to 90°. Formally,

$$\theta_n = \begin{cases} a \times \theta_c & 0 \leq \theta_c \leq 45 \\ a \times (90 - \theta_c) & 45 \leq \theta_c \leq 90 \end{cases}. \tag{3.1}$$

Figure 3.5 visualizes θ_n (water-plane tilt) based on θ_c (container tilt) for different choices of the coefficient a. The water-plane normal ($\overrightarrow{\boldsymbol{n}} = (0, 1, 0)$ in world coordinates) is rotated by $\boldsymbol{\theta_n} = \theta_n \overrightarrow{\mathbf{e}}$, where $\overrightarrow{\mathbf{e}}$ is the axis of container's rotation. Figure 3.6 depicts an example of the container and water-plane normal rotations.

Figure 3.7 illustrates the water's natural stationary state ($a = 0$) and its anomalous states with different anomaly factors (a = -1, -0.5, 0.5, and 1) for various container rotations. As shown in Figure 3.7, when a is positive, the angle between the liquid-plane normal and ($\overrightarrow{\mathbf{y}_c}$) is smaller than it must be in order to keep the liquid-plane horizontal, in other words, the liquid-plane "under-rotates" relative to the container and remains closer to being parallel to

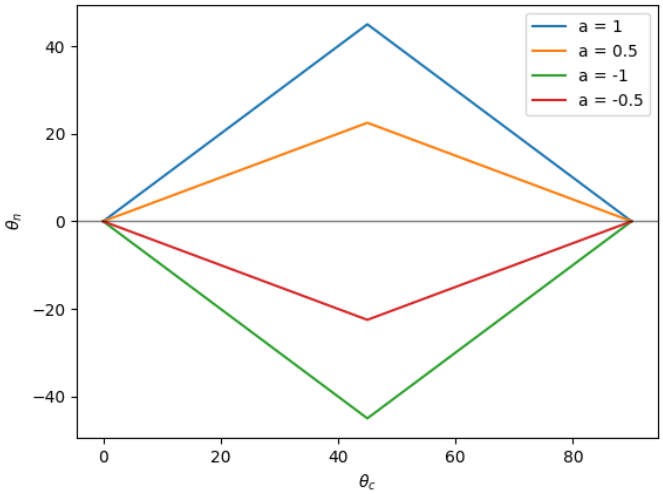


Figure 3.5: Liquid normal rotation angle (θ_n) based on the container's rotation angle (θ_c) for different a values.

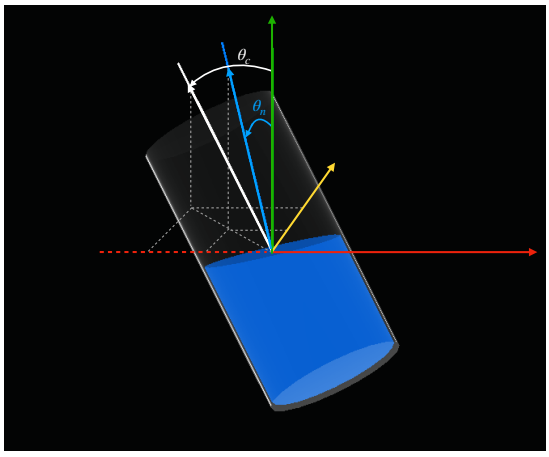


Figure 3.6: The liquid plane's normal (light blue vector) is rotated $\theta_n = a\theta_c$ degrees about the same axis as the container's rotation (θ_c degrees from upright). The white vector represents the container's axis of symmetry. The red, green, and yellow vectors represent world's x, y, and z coordinates, respectively.

the bottom of the container. On the other hand, when a is negative, the angle between the liquid-plane normal and ($\overrightarrow{\mathbf{y}_c}$) is larger than it must be, which can be seen as the liquid-plane "over-rotating" towards the rim of the container. Note that when $a = 0$, the liquid-plane is not altered, and the liquid's stationary state remains horizontal.

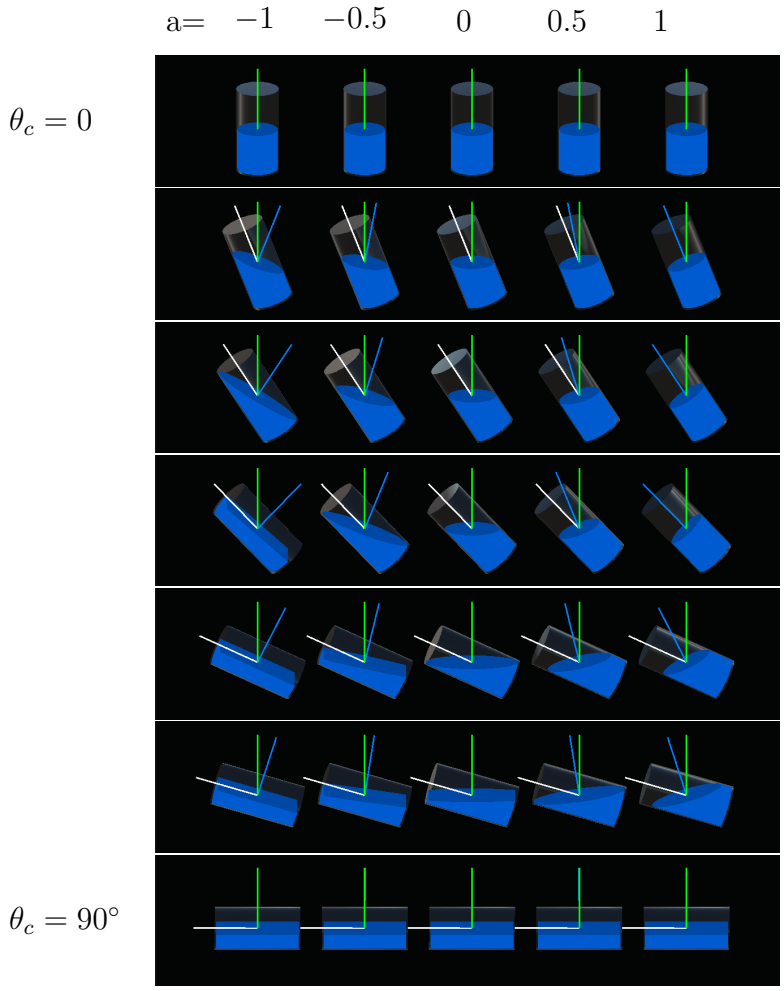


Figure 3.7: Water surface stationary state with anomaly factor -1 and -0.5 ("over-rotation"), 0 ("natural"), and 0.5 and 1 ("under-rotation") from left to right. The container is rotated from 0° to 90° about the world z-axis from top to bottom. The water-plane normal ($\overrightarrow{\boldsymbol{n}}$), container's axis of symmetry ($\overrightarrow{\boldsymbol{y}_c}$), and world y-axis are shown as blue, white, and green lines.

To summarize, our alterations of the water-plane had the following properties:

1. Water was horizontal for vertical and horizontal container orientation.

2. When the container was tilted, the water-plane would tilt about the same axis and proportional to the container's tilt.

3. Water would touch the rim of the container at a container tilt between 0 and 90 degrees (i.e. it poured out at some point)

4. Water would touch the rim of the container in the correct spot but at a wrong container tilt (when the container was too flat (over-rotating) or too upright (under-rotating)).

Finally, as the container was moved and rotated, the altered water-plane oscillated about its stationary state as discussed in section 2.2.2. This could be viewed as applying the rotation $\theta_n \overrightarrow{\mathbf{e}}$ to the oscillating water-plane normal at each time frame. Altering the water-plane normal as we did is slightly different from changing the world's gravity direction. If we were to change the gravity direction and use our oscillation implementation, we would have had to calculate the component of the container's acceleration perpendicular to the gravity and use that force to calculate the surface oscillations. However, we believe the difference is negligible, and our method satisfies our goal of investigating if "wrong" physics feels wrong.

3.3.2 AR-WLT Procedure

For the AR-WLT task, the recruited participants wore a HoloLens2 device connected to an Xbox controller. Before starting the task, the device was calibrated for each participant's eyes. The AR-WLT consisted of 20 trial conditions, each of which was repeated five times. The details of each condition will be discussed in section 3.3.2.

In each trial, the subjects had to interact with two half-full virtual containers and use the Xbox controller to choose which container looked more realistic and submit their answers. The participants could rest the Xbox controller on their laps while interacting with the containers, or they could hold the controller in one hand and interact with the containers

with the other hand. The participants could interact with both containers simultaneously; however, interaction data analysis showed that participants manipulated the containers one at a time. Figure 3.8 shows the two container's tilts in all frames of two different trials for a participant. We used such plots to analyze participants' interactions with the containers. As the figure shows, the orange and blue lines did not change simultaneously, meaning the participant interacted with the containers individually.

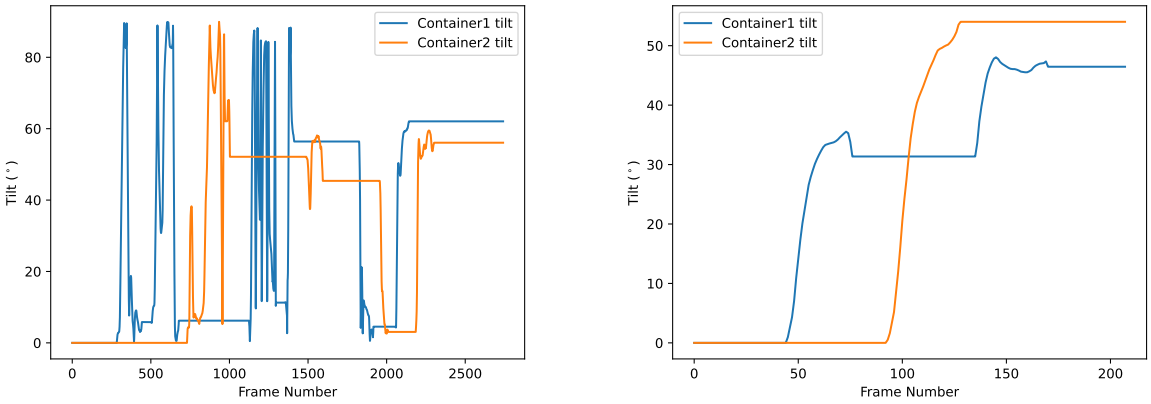


Figure 3.8: Sample of interaction data plots for a participant's first and last trials (left and right, respectively). The plots show that the two containers were not moved simultaneously.

The colour of the water was different for the two containers for the participants to distinguish between them. Apart from having done the prescreening (described in section 3.2), no information was given to the subjects about the horizontality principle. The participants were told to assume the two containers were half full of water and to move and rotate the containers for each trial. It was also hinted to them to pay attention to how the water surface orientation reacted to the movement and rotation of the container. The participants were asked to choose the simulation that felt or looked more natural. At the beginning of each trial, the Xbox controller's inputs were disabled until the participants had interacted with both containers. The experiment explanations were orally given to the participants before they put on the HoloLens2 and saw the trials. The task description was also provided on an

instruction page in the Hololens2 application before the beginning of the trials.

Figure 3.9 shows a single trial from a participant's view. As illustrated in figure 3.9, participants could use virtual and real-world horizontal cues (e.g. the virtual instruction box and the actual table surfaces) to determine if the surface of the water was horizontal.

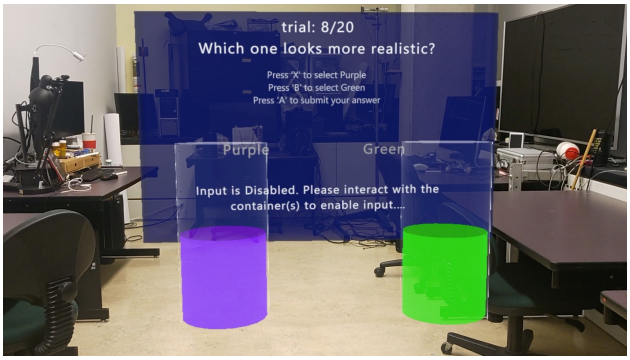


Figure 3.9: An AR-WLT trial from the participant's view.

Lastly, figure 3.10 shows a participant doing the AR-WLT.

Figure 3.10: A participant doing the AR-WLT.

3.3.3 Details of the AR-WLT Trials

In each trial, participants saw two containers containing green and purple coloured water. Each trial setting was defined as a combination of three values a_1, a_2, and $c \in \{True, False\}$. Values a_1 and a_2 were the anomaly coefficient a (discussed in section 3.3.1) determining the water-plane normal for the two containers' water, and c determined if the containers were covered when the participants interacted with them. In the covered trials, a white box appeared around a container as soon as the participant touched it and disappeared when it was released. The participant could still see a slight water movement after releasing the container; however, the liquid's reaction to the container movements was not observable in the covered trials. Figure 3.11 shows interaction with the containers for two similar trials with $c = True$ and $c = False$.

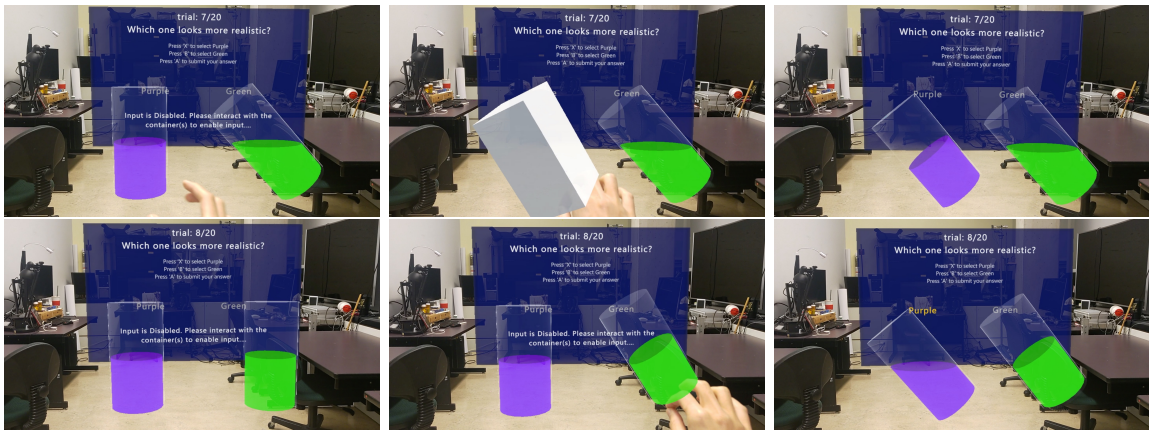


Figure 3.11: Interaction with the containers in a covered ($c = True$) and an uncovered ($c = False$) trial (top and bottom figures, respectively). The left, middle, and right pictures show the participant approaching, grabbing and manipulating, and releasing the container. In the covered trials, the participants could not see a container while they were interacting with it.

We used ten combinations of a_1 and a_2 with both $c = True$ and $c = False$, resulting in 20 conditions. Table 3.1 summarizes the a_1 and a_2 values used for the trials.

As shown in table 3.1, we used seven a_1 and a_2 combinations in which one simulation

Table 3.1: AR-WLT a_1 and a_2 combinations.

(a) Correct vs under-rotating.

a_1	a_2
0	0.3
0	0.5
0	0.7
0	1

(b) Correct vs over-rotating.

a_1	a_2
0	-0.3
0	-0.5
0	-0.7

(c) Two incorrect settings.

a_1	a_2
-0.3	0.3
0.3	0.5
-0.3	-0.5

was correct (i.e. the liquid-plane remained horizontal) and three combinations where neither of the simulations was correct. The trials with two incorrect simulations tested whether the participants found the simulation in which the liquid remained closer to horizontal more realistic (0.3 vs 0.5 and -0.3 vs -0.5 settings). Also, we were interested in knowing whether over-rotating and under-rotating simulations felt more natural to the participants (the -0.3 vs 0.3 setting). As mentioned before, each condition was repeated five times resulting in a total of 100 trials, which were shown to participants in random order. In each trial, a_1 and a_2 were randomly assigned to the green and purple waters.

Data and Evaluation

To evaluate the participants' performance on the AR-WLT, we calculated two scores for each participant: AR-WLT score in correct trials (referred to as the **AR-WLT-z** score) and AR-WLT score in all trials (referred to as the **AR-WLT-a** score). AR-WLT-z considers only the trials in which one of the two simulations was correct (i.e. $a_1 = 0$, tables 3.1a and 3.1b) and shows in what percentage of those trials the participant chose the "correct" simulation (where $a = 0$). For the AR-WLT-a score, we included all trial settings except the ($a_1 = 0.03$, $a_2 = -0.03$) condition, in which the water plane's maximum tilt from horizontal was the same for both containers and only the direction of the tilt (over-rotation vs under-rotation) was different. The score determines in what percent of all the trials, except the excluded

one, the participant selected the "better" simulation. Here, the "better" simulation is the one in which the water's stationary orientation remains closer to horizontal. In other words, the absolute value of the anomaly coefficient is smaller. For each trial, we saved the trial conditions (a_1, a_2, c, and the colour assignment), and the trial result (the participant's choice), which are necessary for calculating the participants' scores.

Additionally, from the start to the end of each trial, we recorded the position and rotation of the participant's head and the two containers in each frame to analyze the participant's interactions with the containers and find out if the position and orientation of the participants' head and the two containers affected their success in trials. Each position was recorded as **x**; **y**; and **z** coordinates. The rotation of the container was saved as an Euler angle representation. To analyse the effect of the container's rotation, we summarized the rotation as two angle values, θ and ϕ. As the cylinder is symmetrical, we assume the container's local y-axis is always pointing upward in the world coordinates. θ equals to θ_c (defined in previous section) and represents the container's tilt from upright and ϕ represents the direction of the tilt. Specifically, the container's orientation can be determined by applying a θ degrees rotation about the world's z-axis (forward direction) and then ϕ degrees about the world's y-axis. As in Unity engine world's z-axis points forward, a ϕ value between 180° and 360° means the container was tilted towards the participant, and if $0 \leq \phi \leq 180$, the container was tilted in forward direction. Figure 3.12 shows θ and ϕ for a container. Head rotation θ and ϕ were calculated similar to container's rotation, assuming upright posture.

3.4 The control task

The control task was designed to evaluate the participants' ability to explicitly judge the horizontality of the water surface in the AR environment while wearing the Hololens2 device. We performed the control task to assess and control for errors caused by participants

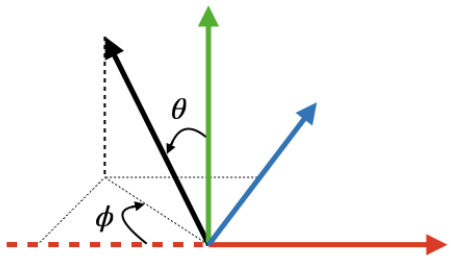


Figure 3.12: θ and ϕ for a container tilted away from the participant. The red, green, and blue arrows are the x, y, z axes, respectively. The black arrow represents the Container's axis of symmetry ($(0, 1, 0)$ in container's local coordinates if it is pointing up, otherwise $(0, -1, 0)$ in container's coordinates).

having difficulty determining the orientation of the water surface in the virtual environment, especially when the container was tilted (orientation illusion).

At the beginning of the AR-WLT, the participants were told there would be a short experiment after the AR-WLT. However, we only included the details of the control task after the participants had finished the AR-WLT to prevent the instructions from affecting their judgements for the AR-WLT. The app automatically showed the descriptions of the control task after AR-WLT trials finished, and the participants could start it right away. However, they could also rest before starting the control task or ask questions if they had any.

In each control task trial, participants were asked to adjust a surface to make it horizontal or parallel to the ground. They could rotate the virtual surface by grabbing it and rotating their hand. When satisfied with their adjustments, they would use the Xbox controller to submit their answer. We used the same liquid inside a container shader used in the AR-WLT (refer to section 2.2.2). However, the container and liquid-plane's positions were fixed, and the liquid surface did not oscillate. Moreover, the surface only rotated about the world's z

axis. Figure 3.13 shows the control task from the participants' view.

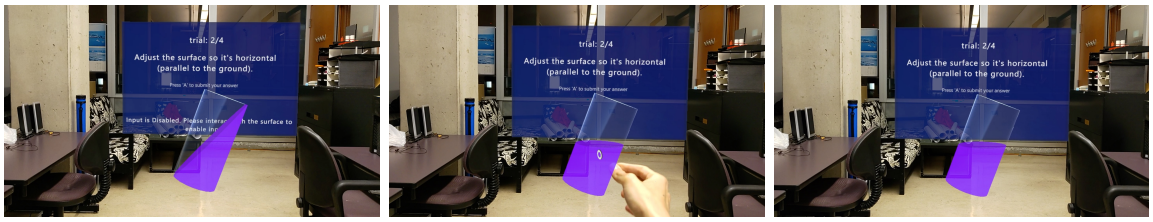


Figure 3.13: The control task: Interaction with the surface from the participants' view. The left picture shows the initial trial condition. In the middle image, the participant is adjusting the surface, and in the right picture, the adjusted surface is being submitted.

Each condition varied in the container rotation and the initial rotation of the plane about the z axis from the upright position. The rotation of the container and plane's normal about the x axis were zero for all conditions. The adjusted plane's normal was saved as the output of each trial.

We tested four different conditions (table 3.2) and repeated each three times, resulting in twelve trials. As mentioned before, Sholl & Liben (1995) showed the orientation illusion is at its maximum around 20° of tilt; hence, the container tilt was set to 20° and −20° in the first three conditions, with the initial plane tilt of 45° and −45° for container tilt of 20° and the initial plane tilt of 45° for container tilt of −20°. We also included the last setting, where the outside frame (container) was upright, to evaluate the participants' perception of horizontal surfaces in the AR environment without being affected by the tilt illusion.

Table 3.2: The control task settings.

container's tilt	**liquid's tilt**
20	45
20	-45
-20	45
0	-45

Data and evaluation

To evaluate performance on the control task, as many participants reported having submitted some trials by mistake, we selected each participant's best effort for each trial setting. In other words, for each condition, we chose the trial where the angle between the adjusted plane normal and the world y axis was minimum as each participant's submitted surface. The reason for the erroneous submissions was that when participants moved their hands towards the controller after having adjusted the plane horizontally, sometimes the HoloLens2 hand-tracking did not identify that they had released the surface, so the surface kept rotating. We defined a participant's **answer** for a condition as the signed angle between their submitted surface and the horizontal surface. Then, we calculated participants' **score** as the absolute value of their answer for each trial setting (lower values correspond to better accuracy). Finally, each participant's **final score** for the control task was calculated as their average score over the four trial settings (lower values correspond to better accuracy).

Chapter 4

Results

4.1 Introduction

In this chapter, the results of the prescreening, control, and AR-WLT tasks are discussed. Section 4.2 provides the participant's score distribution in the prescreening study and discusses how we chose the low and high-scoring groups. Section 4.3 includes information about the invited high and low-scoring individuals participating in the AR-WLT. As mentioned in section 3.4, we had a control task to evaluate participants' ability to judge the horizontality of the water surface in the AR environment. In section 4.4, we discuss the results of the control task as it was fundamentally important for the participants to understand and estimate the horizontality of surfaces in the AR environment to score high on the AR-WLT. Then, we analyze the participants' overall performance on the AR-WLT task and how it related to their performance on the prescreening and control tasks in sections 4.5 and 4.6.

The "Stats Models" (Seabold & Perktold, 2010) Python package was used for the regression analyses and creation of quartile-quartile plots. Unless otherwise stated, all other statistical tests and examinations were done using the "SciPy" Python package. The visualizations were created using Matplotlib and Pandas Python packages. We used a 0.05 level

of significance in all our statistical analyses.

4.2 Prescreening

In total, 120 undergraduate students participated in the online prescreening test (discussed in section 3.2). We removed two participants' data from our analyses because of invalid data (One participant had drawn one point on each container and the data of one trial was missing for the other participant). Twenty-six participants identified as male, and the remaining ninety-two identified as female. None of the participants chose other gender options (which were "non-binary", "transgender", "intersex", "other", and "prefer not to say"). The average age of the participants was 23.25, with the majority (75%) being between 19 and 24 years old. Seventy-four participants used devices with a medium screen size (between 300 and 750 pixels), and the remaining forty-four devices had a large screen (larger than 750 pixels). Figure 4.1 summarizes the above information.

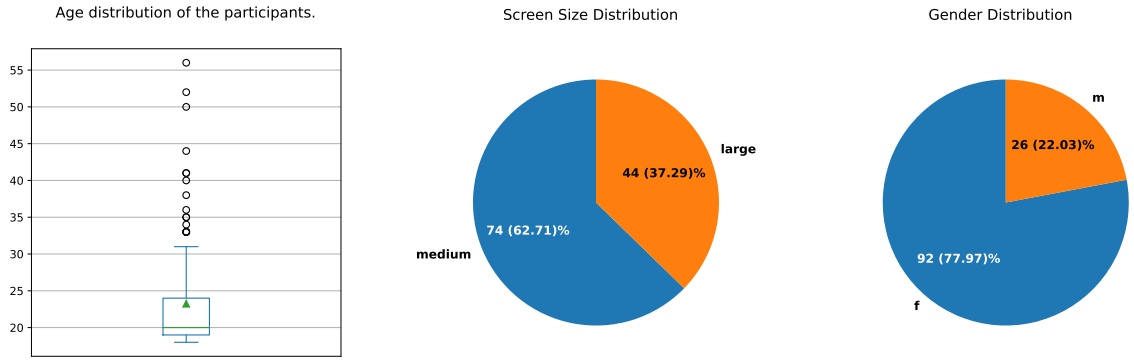


Figure 4.1: The age, gender, and screen size distribution for 118 participants.
The boxplot in this and subsequent figures includes lower and upper quartiles (box), the median (green line), and the mean (green triangle). The whiskers are extended $1.5 \times IRQ$ below and above the lower and upper quartiles, respectively. Data points higher or lower than the whiskers are considered outliers and depicted as circles.

We calculated an individual's score as the number of successful trials (discussed in sec-

tion 3.2) out of the total sixteen trials with a threshold of 15° for success. The remainder of this section summarizes the sensitivity to this choice of the acceptance threshold for success and the impact of age, screen size, and gender on the participants' performance on the prescreening task.

4.2.1 Preliminary Results and Choosing the Acceptance Threshold

As mentioned in section 3.2, we calculated the score of each participant using different acceptance thresholds ($th = 5°$, $10°$, and $15°$), which determines the maximum accepted angle between the participant's answer and the horizontal line classified as a correct response. The mean and standard deviation of the scores with different thresholds are provided in Table 4.1, and Figure 4.2 shows the score distribution with the three thresholds.

Table 4.1: The mean and standard deviation of the participants' scores in the prescreening test.

Threshold	**Mean Score**	**SD**
5	4.56	4.75
10	6.14	5.86
15	7.12	6.29

We set the acceptance threshold to 15° for a more balanced score distribution. As discussed in section 1.2, the acceptance threshold used by other researchers was typically 5° or 10°. The percentage of high and low-scoring participants varied across different studies and differed from our results. More specifically, compared to most previous research, a lower proportion of participants scored performed accurately in our experiment (1%, 10%, and 15%, with thresholds 5°, 10°, and 15°, respectively). Besides the difference in the populations, the digital and online nature of our study might have caused a higher failure rate; also, our method to measure the answer's tilt varies from the previous work. So, by raising

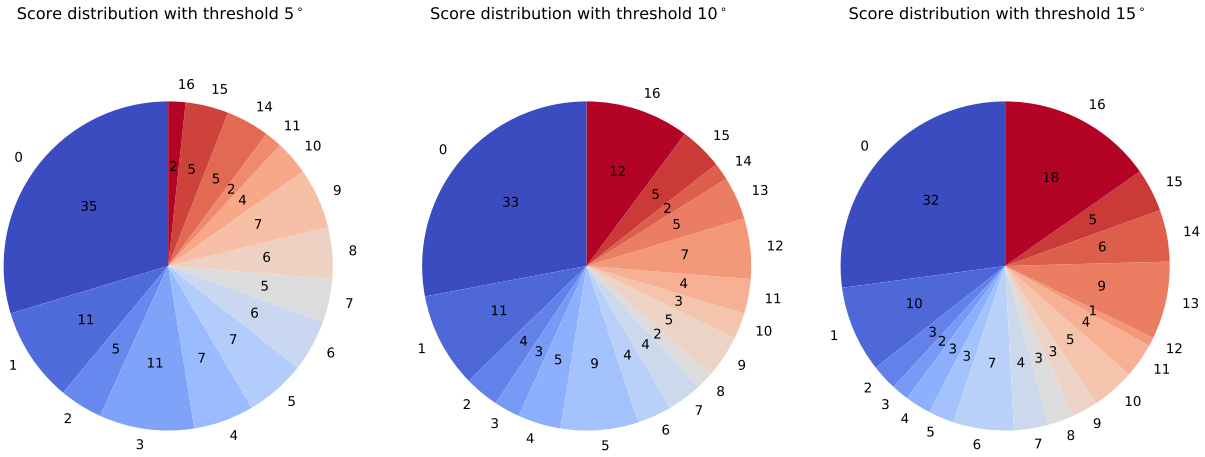


Figure 4.2: The score distribution of 118 participants for the online WLT with acceptance thresholds 5°, 10°, and 15° from left to right. The number outside of each slice represents the score and the number inside each slice is the number of participants with the corresponding score.

the acceptance threshold, we allowed more room for error on each trial, making it easier to score well. The increased acceptance threshold did not significantly affect the number of participants who scored 0 or 1; only 5 participants' scores rose above the score limit of 1 when the threshold was increased from 5° to 15°. However, changing the threshold changed the high-scoring side of the distribution more drastically, favouring individuals who got more tasks approximately correct over the ones who got fewer tasks correct with higher accuracy. We found a strong correlation ($R^2 = 0.94$) between the number of successful trials with acceptance threshold 15° and the participants' answers' average absolute tilt from horizontal (which is an alternative measure of performance in WLT as discussed in section 1.2). Section A.1.1 discusses the effect of our choice of acceptance threshold in more detail.

For the AR-WLT, we invited participants who scored 13 or above as high-scoring and those who scored 1 and 0 as low-scoring, resulting in a total of 80 (42 low-scoring and 38 high-scoring) individuals invited. In total, 35 invited individuals participated in the AR-WLT, whose results will be discussed in section 4.5.

4.2.2 Tilt Illusion

Comparing the low-scoring participants' answer tilts relative to the container tilt for each puzzle showed that most of the low-scoring participants drew lines that were, on average, within 15° of the container tilt (i.e. parallel to the bottom of the container). Figure 4.3 shows the five-point summary of the average absolute value of the participants' answer tilts relative to horizontal and the container tilt for high and low-scoring participants ($N = 38$ and $N = 42$, respectively).

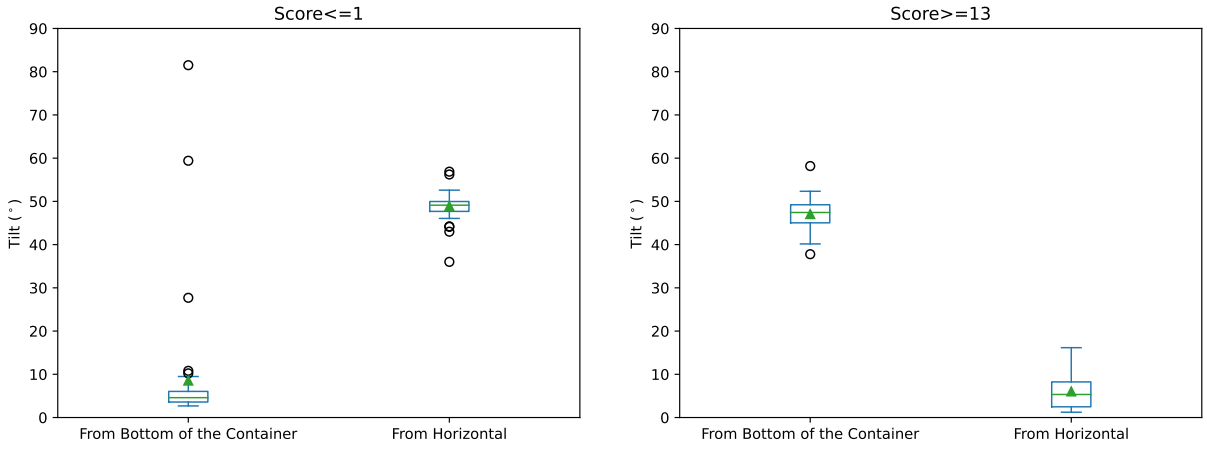


Figure 4.3: The five point summary of the average absolute tilt from horizontal and bottom of the container for low-scoring and high-scoring participants (left and right)

Although similar to the Tilt Illusion (discussed in section 1.2.4), the low-scoring participants' lines tilted to the same direction as the container's tilt, the magnitude of their error was much higher than the Tilt Illusion error (which is reported to be less than 10° from the horizontal), so we believe the reason for their error is not the tilt illusion, but a misconception about how water behaves. In fact, except for 3 individuals (whose answers are provided in section A.1.2), the low-scoring participants' answers were, on average, within 15° of the container tilt. Section A.1.3 analyzes tilt illusion in more detail.

4.2.3 Learning

We compared the number of successful trials in the first and second half of the prescreening task. A one-sided related sample's t-test was performed to compare the number of successful trials in the first and second eight puzzles ($M(118) = 3.26$, $std = 3.12$ and $M(118) = 3.86$, $std = 3.34$, respectively). The results showed that the participants' performance was significantly better in the second half ($t(117) = -4.30$, $p < .001$) although the difference was less than one correct answer on average.

The same analysis was done separately for participants whose scores were below and above 8. We evaluated below and above eight scoring groups instead of low and high-scoring groups because their score ranges were not similar, affecting their performance differences in the first and second halves. In other words, because the low-scoring participants' scores were 0 or 1, the maximum possible difference was 1, while for high-scoring participants, it was higher. The results showed a significant difference in the second half versus first half performance of both groups (above 8 $t(51) = 4.12$, $p < .001$) and below 8 $t(64) = 2.21$, $p = 0.01$). We observed that the average late vs early difference for the above-eight group ($M(51) = 0.95$, $SD = 1.65$) was higher than the below-eight group ($M(64) = 0.35$, $SD = 1.27$).

4.2.4 The Effect of Covariates on the Prescreening Scores

We used multiple linear regression with ordinary least squares (OLS) error to analyze the effect of age, gender, and screen size on prescreening scores. The results showed the predictors explained 8.3% of the variance ($R^2 = 0.083$, $F(3, 114) = 3.451$, $p = .019$). The analysis indicated that gender and the screen size of the device used to participate in the prescreening affected the participants' performance; however, the screen size effect's significance level was borderline ($\beta = 2.34$, $p = .048$). Gender significantly affected the prescreening score ($\beta = 3.01$, $p = .030$). No significant effect was found for age ($\beta = -0.02$, $p = .841$).

Additionally, the screen size and gender could only describe a small proportion (8.3%) of the variance in the score, which means other individual differences were the main factor affecting one's performance in the WLT. The details of the regression analysis are summarized in Table 4.2.

Table 4.2: The linear regression results for analysing the effect of age, gender, and screen size on the participants performance in the prescreening WLT.

Dep. Variable:	score	**R-squared:**	0.083
Model:	OLS	**Adj. R-squared:**	0.059
Method:	Least Squares	**F-statistic:**	3.451
No. Observations:	118	**Prob (F-statistic):**	0.0190
Df Residuals:	114	**Log-Likelihood:**	-378.74
Df Model:	3	**AIC:**	765.5
Covariance Type:	nonrobust	**BIC:**	776.6

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	5.9299	1.907	3.110	0.002	2.153	9.707
age	-0.0150	0.075	-0.201	0.841	-0.163	0.133
gender_m	3.0126	1.368	2.202	0.030	0.303	5.723
screen_size_600	2.3407	1.173	1.995	0.048	0.017	4.665

The effect of gender and screen-size on the prescreening score is visualized in figure 4.4.

4.2.5 The Effect of the Shape and Tilt of The Containers

We calculated the participants' success rate for each trial as the number of successful lines drawn divided by the total number of participants (118). The beer bottle rotated to 270° (Figure 4.5) had the highest success rate of 0.65, and the lowest success rate was 0.34 for the beer bottle rotated 135° and the dish-soap-container rotated 115° (Figure 4.6). The average participants' success rate for each puzzle is included in section A.1.4.

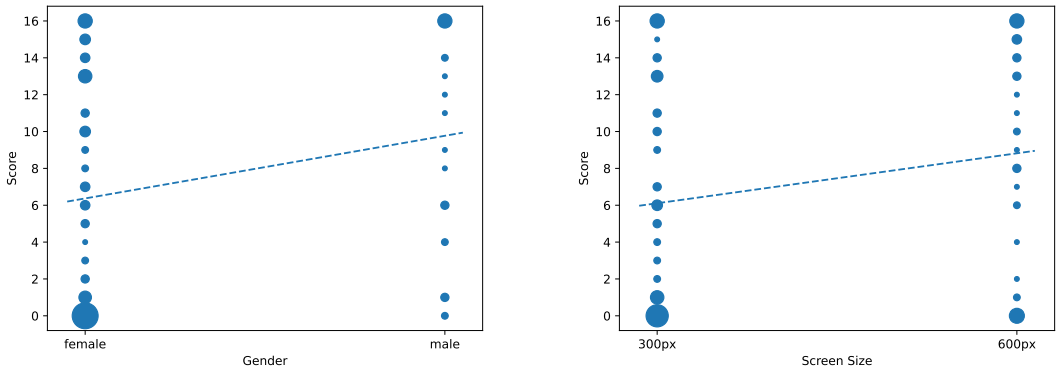


Figure 4.4: The relationship between the gender (left) and screen-size (right) on prescreening performance. The dot size is proportional to the number of data points.

Figure 4.5: The prescreening trial with the highest success rate.

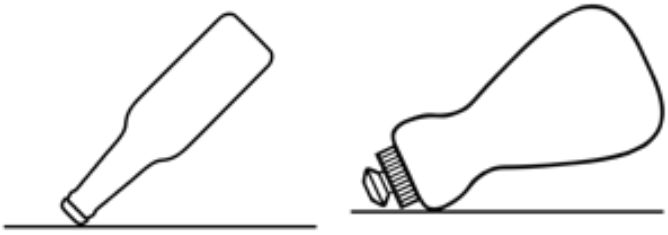


Figure 4.6: The prescreening trials with the lowest success rates.

4.2.6 Discussion

As in previous research on WLT, we found a gender effect in our prescreening task. Moreover, we also found that the screen size of participants' devices affected their performance. However, gender and screen size accounted for a small proportion of variance in the performance suggesting individual differences were the main factor.

Additionally, we found that most low-scoring and high-scoring participants were consistent in different tasks, drawing lines parallel to the bottom of the container and the horizontal, respectively. We could not find patterns in participants whose score was in the

middle. Finally, we found that the tilt illusion affected high-scoring participants' answers, but based on our data, it is most likely that the errors made by low-scoring participants were not attributable to the tilt illusion but were influenced by wrong heuristics.

4.3 AR-WLT Participants

The main goal of our experiment was to test the Water-Level Task in a more realistic and interactive setting. We were interested to know if the misconception about the liquid orientation would be present if, rather than being asked to imagine the water surface's state, the participants could interact with the water and judge the surface's orientation in a more natural setting.

In total, 35 out of 81 high-scoring and low-scoring individuals participated in the AR-WLT, 18 of whom scored 13 or more on the prescreening task (high-scoring group), and the remaining 17 scored 1 or 0 (low-scoring group). Table 4.3 summarizes the prescreening score, age, and the number of male and female participants for the high-scoring and low-scoring groups. The number of high-scoring and low-scoring male and female individuals who participated in the AR-WLT is shown in figure 4.7.

Table 4.3: The low-scoring and high-scoring groups' average score on the prescreening test, age, and number of female and male participants.

	WLT Score		**Age**		**#Participants**	
	mean	**SD**	**mean**	**SD**	**female**	**male**
High-Scoring	14.84	1.25	20.78	3.12	11	7
Low-Scoring	0.24	0.44	25.52	10.71	15	2

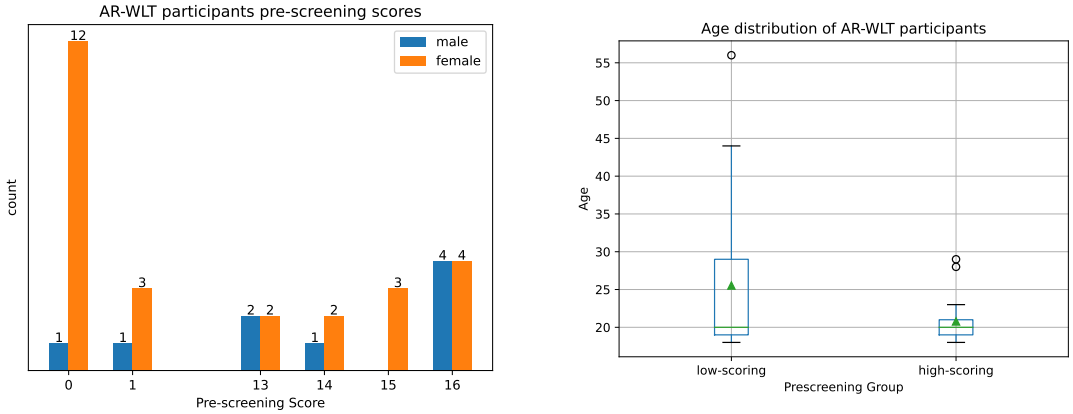


Figure 4.7: The number of male and female participants based on prescreening score (left) and the five-point summary of high-scoring and low-scoring participants' age (right).

4.4 The Control Task

As mentioned in section 3.4, we included the control task to evaluate participants' judgement of the horizontality of surfaces in the AR environment. The study consisted of four trial conditions. We repeated each condition three times and used each participant's best effort (closest to horizontal) among the three repetitions for calculating their **answer** (the signed angle between their best-submitted surface and the flat surface). Each participant's **score** on each condition was calculated as the absolute value of their answer, and each participant's **final score** was calculated as their average score on the four conditions. The final score was 8.58° on average ($SD = 14.83°$) for all participants. Figure 4.8 shows each participant's final score and the final scores summary.

As shown in figure 4.8, one individual aligned the surfaces vertically instead of horizontally (the mean absolute angle between the horizontal surface and the aligned surface over all trials for this participant was 87.21° ($SD = 2.21$)). We excluded this participant from our analyses of the control task results. However, we did not exclude them from the AR-WLT analysis because their control results showed they understood the horizontal and vertical axes in the AR environment well, but they misunderstood the control task.

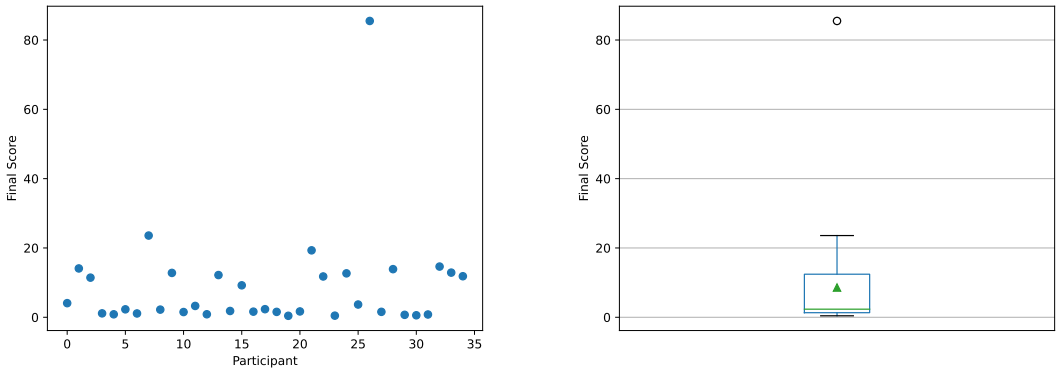


Figure 4.8: The final scores summary (right) and each participant's final score (left) in the control task.

For the remaining 34 participants, the average control task final score was 6.32° (SD = 6.49°), with 94.12% scoring below the 15° threshold used in the prescreening, which means, on average, the participants had a sufficient understanding of the horizontal and vertical surfaces in the AR environment. Figure 4.9 shows the individual final scores and the five-number summary of all final scores with the outlier participant removed.

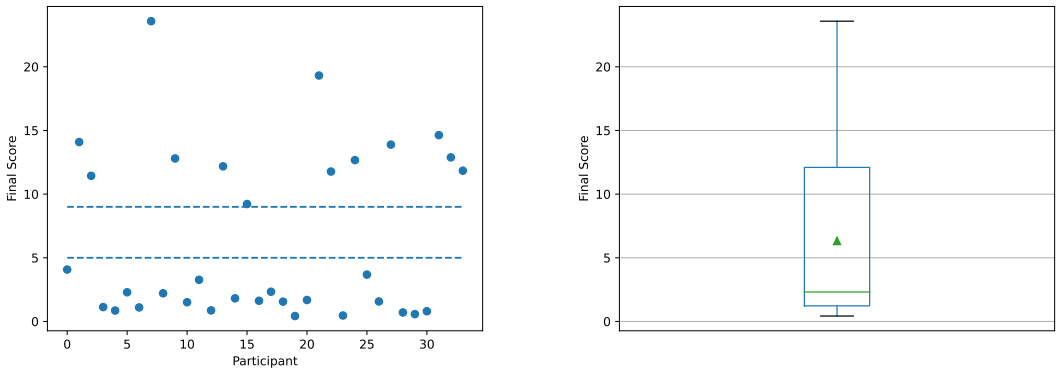


Figure 4.9: The final scores summary (right) and each participant's final score (left) in the control task without the outlier.

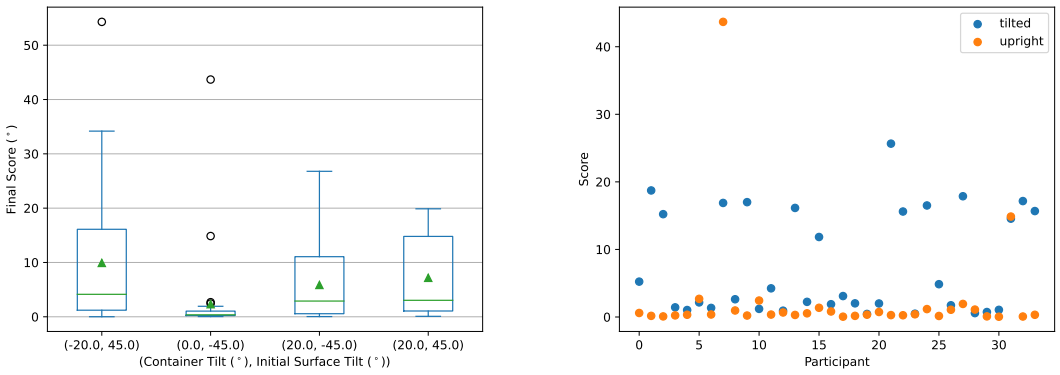
As suggested in figure 4.9, the control task divided the participants into two groups, with 21 participants scoring below 5° ($M(21)$ = 1.65°, SD = 1.03°), and 13 participants

scoring above 9° ($M(13) = 13.87°$, $SD = 3.73°$). As the participants showed a distinction in their performance on the control task, we have included each participant's final score as an independent variable in the AR-WLT task analysis.

The remainder of this section includes an analysis of the participant's performance in the control task's different settings and the correlation between the prescreening and control task performances.

4.4.1 The Effect of Condition on Participant's Performance

Figure 4.10a shows each participant's score in each trial condition. There is a clear distinction between scores on the condition where the container tilt upright (container tilt = 0°) and the conditions with a tilted container. Figure 4.10b shows each participant's average score in the tilted tasks along with their score in the upright task, making the distinction clearer. In fact, except for two participants who scored 14.86° and 43.67°, all other participants scored below 3° in the upright task.



(a) Each trial condition.

(b) Upright vs tilted condition.

Figure 4.10: Comparison of participants' score (unsigned error) on the four control task settings.

To formally compare the upright and tilted conditions, we averaged each participant's

score on the three tilted versions and compared that to their score on the upright one. A Shapiro-Wilk test for normality (SHAPIRO & WILK, 1965) showed both values differed significantly from a normal distribution ($W = 0.293$, $p < .001$ for upright scores and $W = 0.796$, $p < .001$ for average tilted scores). So, we used a Wilcoxon signed-rank test to compare the two paired values. The results showed participants' scores were significantly higher on the tilted tasks ($Mdn = 2.86$) than on the upright task ($Mdn = 0.36$), ($V = 63.0$, $z = -4.009$, $p < 0.001$), which means they performed poorer on the tilted task. Lower performance on the tilted tasks can be attributed to the tilt illusion or the participant's chosen frame of reference on the tilted tasks. Specifically, as shown in figure 4.10b, ten participants' average scores for the tilted conditions were within five degrees of the container tilt (20°), which suggests they aligned the surface parallel to the bottom of the container rather than the ground. Table 4.4 summarizes the average participant performance on each trial setting.

Table 4.4: The median, mean, and standard deviation of participants' scores in each control task condition.

#	Container Tilt	Initial Surface Tilt	Median	Mean	SD
1	20	45	3.03	7.18	7.17
2	20	-45	2.89	5.86	7.12
3	-20	45	4.14	9.92	12.12
4	0	-45	0.36	2.32	7.73

We used Wilcoxon signed-rank test to compare participants' performance on different container tilts and initial surface tilts. Comparing settings 2 and 3 (the two settings are symmetrical about the y axis) shows the participants performed better on average when the container was rotated clockwise ($V = 176.0$, $z = -2.772$, $p = .019$). And comparing settings 1 and 2 shows for the clockwise rotated container, the setting with counter clock-wise initial surface tilt resulted in better performance ($V = 191.0$, $z = -1.820$, $p = .034$).

We used participants' scores on the control trials for the previous analyses, which were the absolute value of their submitted surface's tilt. Taking the absolute values of participants'

answers prevented the negative and positive tilts from cancelling out and gave us a better understanding of the average overall accuracy of participants. However, it did not show the relationship between the container's tilt direction and that of the submitted surfaces. Therefore, we also plotted the distribution of the participants' signed responses for each of the container tilt values. Figure 4.11 shows the distribution of participant's answers for container tilt values of $-20°$ (total trials=34, $M = -3.74$, $SD = 15.29$), $0°$ (total trials=34, $M = -0.90$, $SD = 8.03$), and $20°$ (total trials=68, $M = 2.66$, $SD = 9.31$).

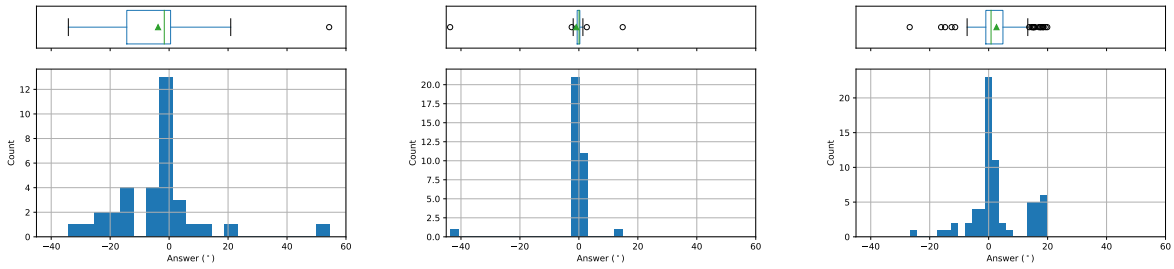


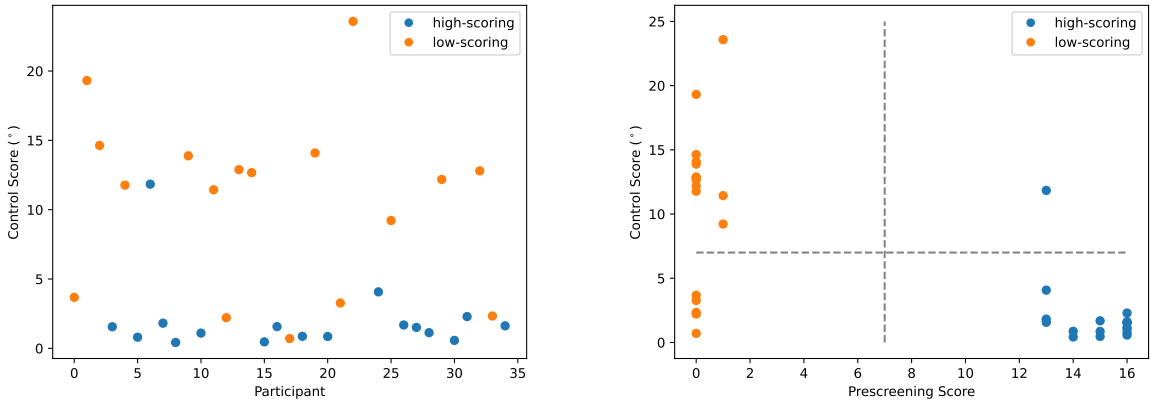
Figure 4.11: Participants' answers five point summary and distribution for control trials with container tilt -20, 0, and 20 from left to right.

As the figure shows, the answers were skewed in the same direction as the container tilt. This observation is the same as tilt Illusion (discussed in section 1.2.4), in which participants perceived a line tilted in the same direction as a frame tilt as horizontal. However, as some participants' final answers were tilted about $20°$ (i.e. the same amount as the container), it is also possible that they misunderstood the task and aligned the surface parallel to the bottom of the container.

Note that the findings about the effect of the control task conditions and participant performance discussed in this section are preliminary and need to be further analyzed with more data (i.e. more repetitions per task and more participants). We used this task to ascertain whether an average participant had an adequate understanding of horizontal surfaces in the AR environment, not to investigate the effect of different settings.

4.4.2 The Correlation Between the Prescreening and Control Tasks

This section investigates the correlation between the control and prescreening tasks. Figure 4.12a plots each participant's control score, and each dot is coloured based on the participant's prescreening group. The plot shows that, except for one participant, the individuals who did not perform well on the control task were among the low-scoring group. The control score is plotted against the prescreening score in figure 4.12b.



(a) Control scores for high and low scoring groups

(b) Control score as a function of prescreening score.

Figure 4.12: The relationship between control and prescreening tasks.

Figure 4.12 shows that control and prescreening scores divide participants in four distinct groups, including high- and low-scoring (prescreening) individuals who did poorly and well on the control task. Table 4.5 shows the contingency table for the two variables. As one value in the contingency table was less than 5, we used the Fischer exact test to determine the significance of the association between the two tests. The results confirmed that the association between the two tests was significant ($p < .001$).

Table 4.5: The contingency table of prescreening and control tasks.

	high-scoring	**low-scoring**	TOTAL
below 5°	16	5	21
above 9°	1	12	13
TOTAL	17	17	34

4.4.3 Discussion

The Control task results showed participants could identify horizontal surfaces in the AR environment. However, many participants' judgements were affected by a tilted surrounding frame. The effect could be attributed to the Tilt Illusion (discussed in section 1.2.4); participants' misunderstanding of the task could also cause the error. The correlation between the control task and prescreening task performance could be evidence that one's cognitive style (i.e. field dependence discussed in section 1.2.4) affected their performance in both tasks.

4.5 The AR-WLT

The main goal of this study was to examine the participant's understanding of water level. The AR-WLT task examined if the participants could easily distinguish natural from unnatural orientation of water when interacting with containers. An immediately interesting observation was that, despite our prediction, the participants did not easily recognize the difference between the two simulations, and none of the inconsistent simulations immediately seemed less natural to them. When asked if one of the two simulations felt or looked obviously less realistic than the other, all participants said the difference in the simulations was not immediately noticeable. It took the participants a few trials to realize what the difference was.

We calculated the participants' scores as the percentage of their correct answers on trials

with one correct simulation (**AR-WLT-z** score) and all trials except the ($a1 = -0.3$, $a2 = 0.3$) setting (**AR-WLT-a** score). The average participant AR-WLT-z score was 68.58 ($SD = 17.54$) and for AR-WLT-a it was 67.81 ($SD = 17.17$). Figure 4.13 shows AR-WLT-z and AR-WLT-a scores distribution and five-number summary for all participants.

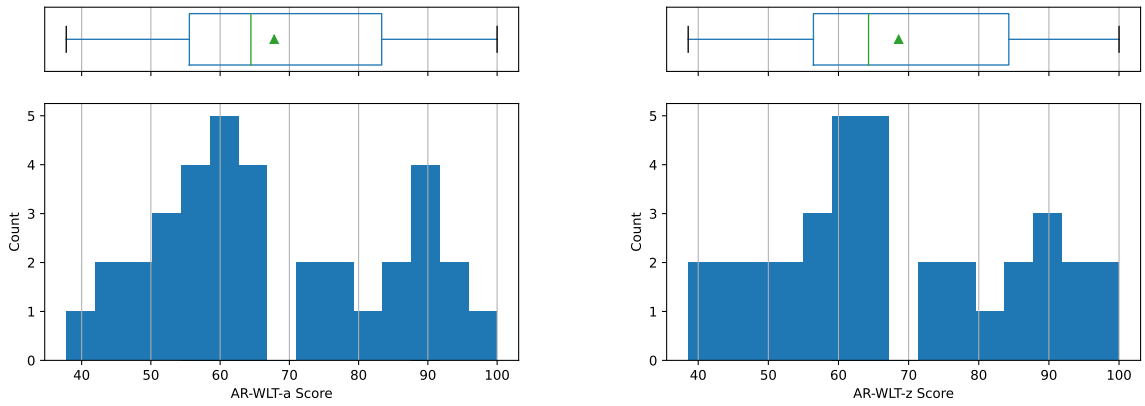


Figure 4.13: AR-WLT-a (left) and AR-WLT-z (right) five-number summary and distribution.

The average score of the AR-WLT shows that the task was not too obvious for the average participant as scores are not near ceiling performance. Also, it can be observed from the distribution histograms (figure 4.13) that the AR-WLT-z and AR-WLT-a scores seem to divide the participants into two groups (with a threshold of 70%). However, the clustering is relatively weak as can be observed from the scatter plots of individual scores (figure 4.14).

Lastly, we ensured that, on average, participants' choices differed significantly from a random choice. We included all trial settings except for ($a1 = -0.3$, $a2 = 0.3$) for this analysis. We conducted individual binomial tests (provided in section A.2.1). On average, the 95% interval for proportion estimates lower-bound was ($M(35) = 0.58$, $SD = 0.18$) and the upper bound was ($M(35) = 0.76$, $SD = 0.14$). We concluded that the average participants performed significantly more accurately than a random choice.

So far, we have examined the overall performance on the AR-WLT and showed that although the participants were more likely to choose the correct answers, on average, the

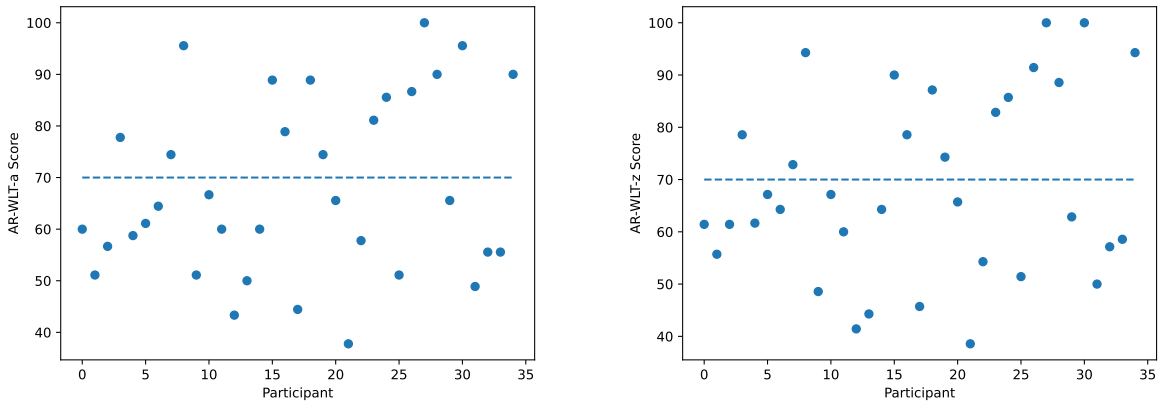


Figure 4.14: AR-WLT-a (left) and AR-WLT-z (right) scores for each participant.

correct answer was not completely obvious. The remaining of this section will assess how different test settings affected performance and how the AR-WLT related to the control and prescreening tasks.

4.5.1 An Analysis of Different AR-WLT Trial Settings

Each AR-WLT condition consisted a combination of three independent variables, $a1$; $a2$; and *covered*, and was repeated 5 times (described in detail in section 3.3.2). In each repetition, a_1 and a_2 were randomly assigned to the purple and green water simulations. In this section, first, we investigate whether the cover and colour affected the participants' choice and the success of the trials. Then, we assess how $a1$ and $a2$ affected the difficulty of trials. Finally, we will analyze the ($a1 = -0.3$, $a2 = 0.3$) trial setting to determine if there was a tendency toward choosing the over-rotating or under-rotating water simulations.

The Effect of Cover

As discussed before, for each $a1$ and $a2$ combination, we added five covered trials and five trials with no cover. The cover appeared when the participants interacted with the containers

to begin manipulation and disappeared when the container was released. We added the covered trials to reduce the visibility of the water movement and to determine whether the participants could successfully perform the WLT without seeing the interaction, only the final orientation. In this section, we analyze the effect of cover on trial success. Because we are analyzing trial success, we again exclude the condition with ($a1 = -0.3$, $a2 = 0.3$).

To examine the effect of the cover for each participant and trial setting ($a1$, $a2$, and *covered*), we calculated the proportion of successful trials (number of successful trials range from 0 to 5) out of five repeats as a participant's success rate for the condition (a number between 0 and 1 with 0.2 increments). Then, we compared the average success rate of each participant in covered and uncovered conditions. The average success rate of participants for the covered test was 0.677 ($SD = 0.174$), and for the trials with no cover, it was 0.678 ($SD = 0.179$). The mean and standard deviations are reported with three decimal points because the means were equal up to two decimal points.

The equality of mean and variance of success rate with and without a cover suggested that the cover did not have a meaningful effect on performance. Figure 4.15 shows the distribution of the differences for 35 paired data points. As shown in figure 4.15, the differences were approximately normal, so we used a paired t-test to evaluate the cover effect. The t-test failed to reject the null hypotheses that the difference in success rate in covered and non-covered trials was significant ($t(34) = -0.130$, $p = .897$). The 95% confidence interval for the difference was $[-0.03, 0.03]$. As the success rate could get discrete values from 0 to 1 with 0.2 increments, the confidence interval meant the participants' success on covered and non-covered trials differed in at most one simulation, so we concluded that any difference was not practically significant. The cover's insignificant effect shows that the water's oscillations and dynamic movement did not measurably affect participants' judgement and did not distract participants' focus from the water surface's stationary orientation.

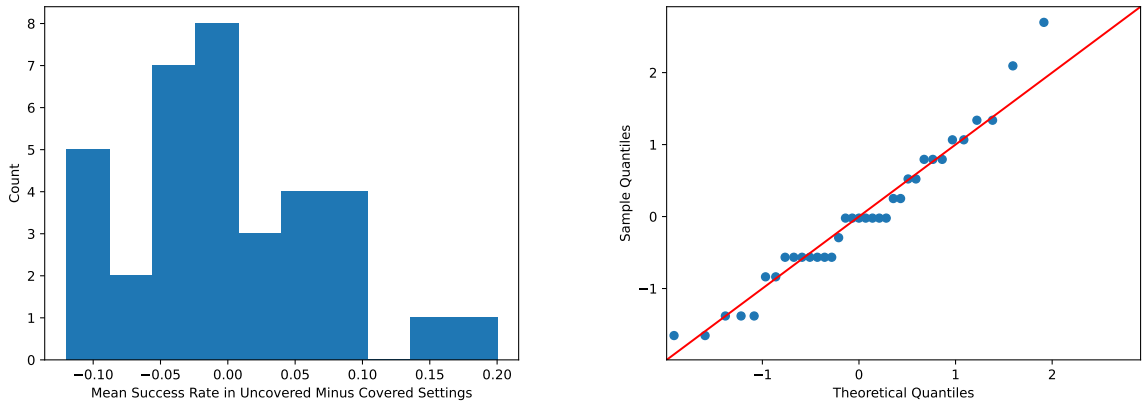


Figure 4.15: The distribution and qq plots of the difference of participants' average success rate for paired covered and non-covered trials.

The Effect of Color

In all trials, the left and right containers were purple and green, respectively, and $a1$ and $a2$ were assigned to them randomly. We compared participants' average success rates in conditions where the correct simulations were purple and green. As in the previous section, we excluded the $(a1 = -0.3, a2 = 0.3)$ setting. The average success rate was $(M(35) = 0.69, SD = 0.17)$ and $(M(35) = 0.73, SD = 0.16)$ in conditions where the correct simulation was green and purple, respectively. Figure 4.16 shows the distribution of the differences for 35 paired data points. Similar to the previous section, a paired-sample t-test showed the participant's average performance was not significantly affected by the colour of the correct simulation $(t(34) = -1.620, p = .0.115, 95\%CI = [-0.08, 0.01])$. Finally, figure 4.16 suggested one participant's success rate was affected by the simulated water's colour (the outlier on the left). We found that this participant chose the green simulation in only 24 out of 100 trials.

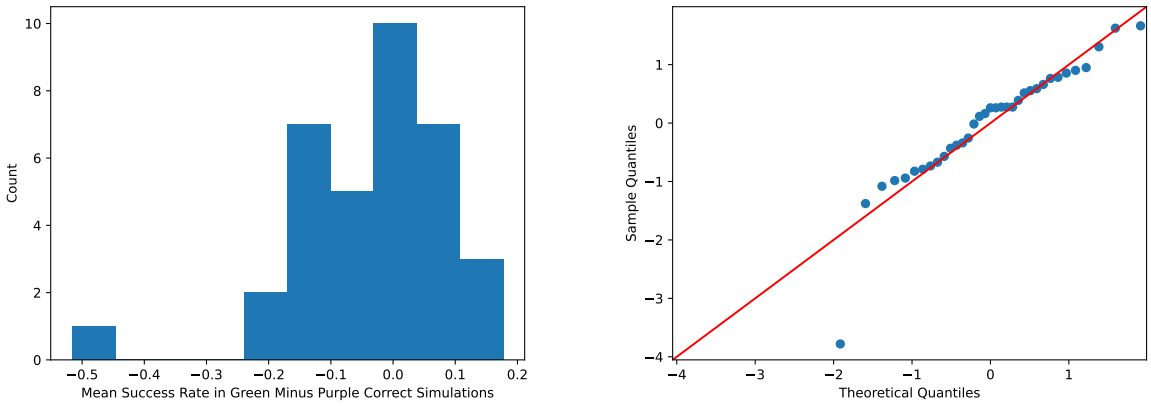


Figure 4.16: The distribution and qq plots of the difference of participants' average success rate in conditions with green and purple correct simulations.

The Average Effect of Anomaly Factors $a1$ and $a2$

In section 4.5.1, we saw that the cover did not significantly affect participants' performance in general, so, in this section, we ignore whether or not each trial was covered. Each combination of $a1$ and $a2$ was repeated ten times for each participant. We calculate $a1$ rate for each participant and $a1$ and $a2$ combination, as the ratio of the trials in which the participant chose the $a1$ simulation to the number of total trials (=10). Note that except for the ($a1 = -0.3$, $a2 = 0.3$) setting, $a1$ rate equals to success rate, as the simulation with $a = a1$ is more realistic (horizontal or closer to horizontal water orientation) for all other conditions. Table 4.6 summarizes the mean $a1$ rate for all participants for each $a1$ and $a2$ combination.

Rows 1-4 of the table show correct vs under-rotating conditions, where the liquid remained close to its initial orientation relative to the bottom of the container. Rows 5-7 are correct vs over-rotating conditions, where, when the container tilts, the liquid moves towards the edge of the container faster than it should. Over and Under rotating settings were discussed in detail in section 3.3.2. Also, a bigger $a2$ shows a bigger orientation error. In row 1, the liquid's still orientation remains parallel to the bottom of the container up to a 45° tilt. The liquid's tilt from horizontal in settings 2, 3, and 4 are the same as rows 5, 6, and 7,

Table 4.6: The $a1$ rate for each $a1$ and $a2$ setting.

#	a1	a2	a1 rate
1	**0**	**1**	0.78
2	**0**	**0.7**	0.75
3	**0**	**0.5**	0.73
4	**0**	**0.3**	0.65
5	**0**	**-0.7**	0.67
6	**0**	**-0.5**	0.63
7	**0**	**-0.3**	0.58
8	**0.3**	**0.5**	0.68
9	**-0.3**	**-0.5**	0.62
10	**-0.3**	**0.3**	0.58

respectively; only the liquid's tilt direction relative to the container's tilt is different.

In the table, we see that, as expected, participants performed more accurately in settings with more exaggerated incorrect simulations (comparing rows 1 to 4 or 5 to 7). Secondly, the data suggest that the under-rotating settings were easier to correctly identify for the participants than the over-rotating ones (compare rows 2, 3, and 4 with 5, 6, and 7, respectively). Figure 4.17 shows the average success rate as a function of $a2$ in settings with $a1 = 0$ (we added the point $a2 = 0$ and with theoretical chance performance of 0.5 to the plotted line). A one sided t-test showed that, on average, participants were significantly more accurate on tasks with $a1 = 0$ and $0 < a2 < 1$ (i.e. rows 2-4, $M(35) = 0.71$, $SD = 0.19$) than tasks with $a1 = 0$ and $-1 < a2 < 0$ (i.e. rows 5-8, $M(35) = 0.63$, $SD = 0.23$), ($t(34) = 2.48$, $p = 0.009$).

The second observation is interesting because most of the low-scoring participants drew lines parallel to the bottom of the container (equal to the most exaggerated under-rotating setting ($a1 = 0.0$, $a2 = 1.0$)) in the prescreening. However, on average participants made fewer mistakes in settings where the incorrect water simulation remained parallel to the bottom of the container. Section 4.5.4 investigates the relationship between low-scoring

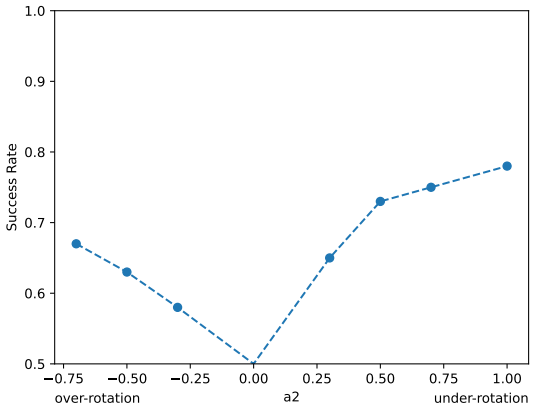


Figure 4.17: The trend of success rate for settings with $a1 = 0$ and different $a2$s.

participants' prescreening and AR-WLT answers in more detail.

In addition, we can see that neither simulation was correct, and both were and under-rotating (row 8), the participants had more difficulty choosing which was more realistic than when one of the simulations was correct (row 3), although the difference in performance was not significant in the two settings ($t(34) = 1.49$, $p = .073$). However, the success rate was almost the same for rows 6 and 9 ($t(34) = 0.25$, $p = .403$), combined with the low success rate in the ($a1 = 0$, $a2 = -0.3$) setting (row 7), the data suggests that $a = -0.3$ looked almost realistic to participants. Lastly, row 10 shows a preference for choosing over-rotating (negative a) compared to the under-rotating (positive a) simulation with the same absolute a value. A one-sided t-test on participants rate of selection for $a1$ and $a2$ showed a significant preference for the negative a value ($t(34) = 1.96$, $p = .029$).

In conclusion, the participants had more difficulty choosing the correct simulations when the incorrect simulation was over-rotating than when it was under-rotating, suggesting the over-rotating simulations looked more realistic to participants than the under-rotating simulations.

4.5.2 Learning

We compared the participants' success rate in the first and second half of the trials for trials other than the ($a1 = -0.3$, $a2 = 0.3$). The results showed their performance slightly improved ($M(35) = 0.67$, $SD = 0.17$ in the first half and $M(35) = 0.69$, $SD = 0.19$ in the second half); however, the difference was not significant ($t(34) = -1.60$, $p = .059$).

4.5.3 AR-WLT Interaction Analysis

For all trials, we saved the rotation and position of the participant's head and the two containers on each frame, as discussed in section 3.3.3. For our analyses, we only considered the last frame of each trial, when the participant made their choice. We examined the final values for position (x, y, and z) and rotation (θ and ϕ) data in each trial. We also calculated the approximate time spent on each trial (in seconds) as each trial's number of frames divided by the average frame rate of 30. In this section, we analyze the interaction data and the correlation between the interaction parameters and the success of the trials.

Section 4.5.3 summarizes the effect of trial number (i.e. the number of trials the participants completed before the current trial) and time spent on the trial. Section 4.5.3 analyses the final rotation and position of the head and the two containers in each trial.

Time and Trial Number

We observed a negative logarithmic relationship between the time spent on each trial and the trial number, which we formally defined using a mixed linear model with $log(\text{trial-number} + 0.0001)$ as a predictor of the time spent on each trial. Table 4.7 summarizes the regression summary[1], and the relationship between the two variables is visualized in figure 4.18. It is

[1]We added the condition ($a1 = 0$, $a2 = 1$) after the first two participants completed the task because we observed that the differences in the already existing trials were not too noticeable for the participants. Thus, we decided to add the most exaggerated case in which the water plane remained parallel to the bottom of the container when it was tilted up to 45°. We express participants' scores as the proportion of their successful

possible that participants spent more time on the first few trials to get used to the task. AS noted in section 4.5.2, this did not seem to improve success rate.

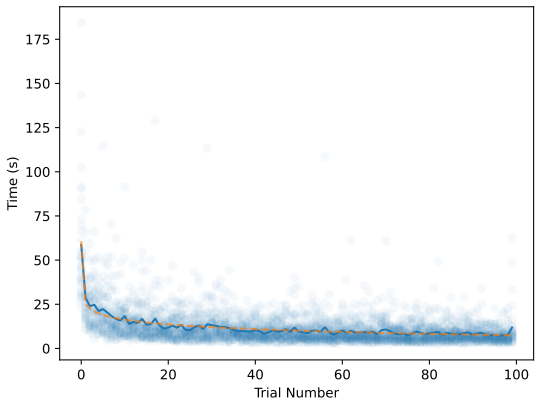


Figure 4.18: The time spent on each trial plotted as a function of the trial number. Each dot represents a single trial (3480 trials in total). The blue line is the average time spent on each trial across all participants. The orange dashed line shows the fitted line.

Table 4.7: The regression results for analysing the relationship between the time spent on each trial and the trial number

Model:	MixedLM	Dependent Variable:	time
No. Observations:	3480	Method:	REML
No. Groups:	35	Scale:	58.6671
Min. group size:	90	Log-Likelihood:	-12081.1031
Max. group size:	100	Converged:	Yes
Mean group size:	99.4		

	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
Intercept	25.160	0.741	33.958	0.000	23.708	26.613
np.log(trial_num + 0.0001)	-3.856	0.082	-46.785	0.000	-4.017	-3.694
id Var	15.724	0.519				

trials to the total trials they did. The absence of the ($a1 = 0$, $a2 = 1$) condition from two participants' data resulted in a total of 3480 trials.

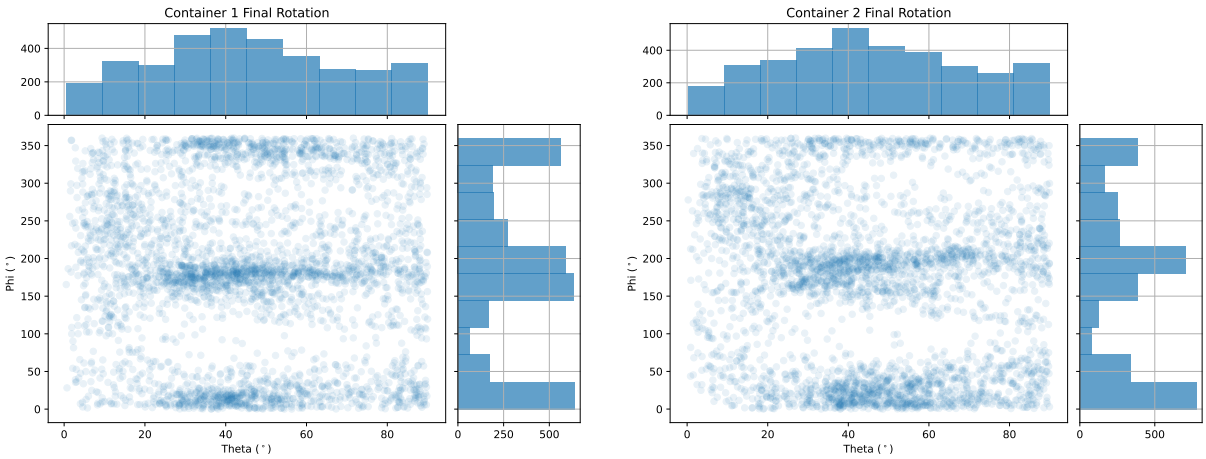


Figure 4.19: The final rotation of the two containers in each trial across all participants and conditions. Each point represents one trial, and darker areas have higher concentration of points.

Rotation and Position of Head and Containers

Figure 4.19 shows a scatter plot of the two containers' final (θ, ϕ) pairs along with separate histograms of θ and ϕ for all trials (3480). As the figure shows, in most trials, the containers were tilted to the right or left ($\phi \approx 0°$ or $360°$ and $\phi \approx 180°$, respectively) in the last frame. Although the amount of tilt (θ) varied between zero and ninety degrees, it was typically between $20°$ and $60°$ relative to level for both containers. This data shows that participants put the two containers in a tilted position without being told to do so, which was the best way to judge which was more realistic.

To investigate whether the participants adjusted the containers to the same tilt, we plotted the θ and ϕ values of container one against container two (figure 4.20). As shown in the figure, the containers tended to be tilted to the same degree for most of the trials (container one and container two's θ values are on the $x = y$ line). We can also see that the two containers were mostly tilted to the same direction, either to right ($\phi = 0°$ or $360°$) or left ($\phi = 180°$).

Moreover, figure 4.21 shows scatter-plots and histograms of the two containers' final

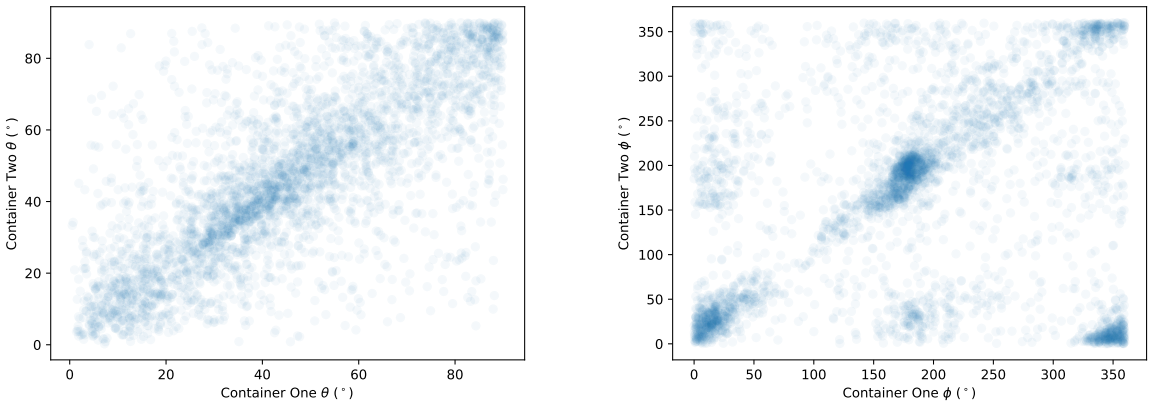


Figure 4.20: The two containers' rotation parameters (θ and ϕ on the right and left, respectively) plotted against each other. Each point represents one trial, and darker areas have higher concentration of points.

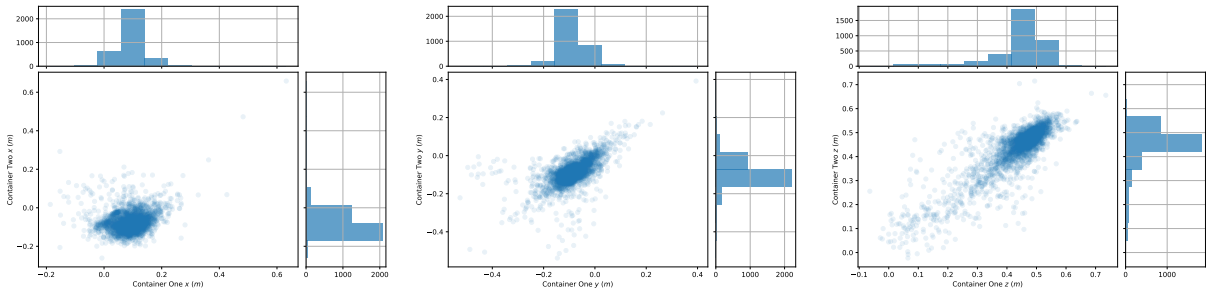


Figure 4.21: The two containers' position on x, y, and z axes (from left to right) plotted against each other. Each point represents one trial, and darker areas have higher concentration of points.

position on x, y, and z axes plotted against each other. We can see that the containers were left in approximately the same location on the y and z axes. And, as the position plot on the x axis shows, the two containers were left beside each other (container one at approximately $x \approx 0.1$ m and container two at $x \approx -0.1$ m).

The participant's head final rotation and position on the horizontal plane (x and z axes in Unity) are plotted in figure 4.22. As the figure shows, in most trials, the participant's head was tilted slightly ($\theta < 15°$) backward ($phi \approx 270°$), which shows they did not need to look down to use the controller. Also, the participant's head remained within 20 centimetres

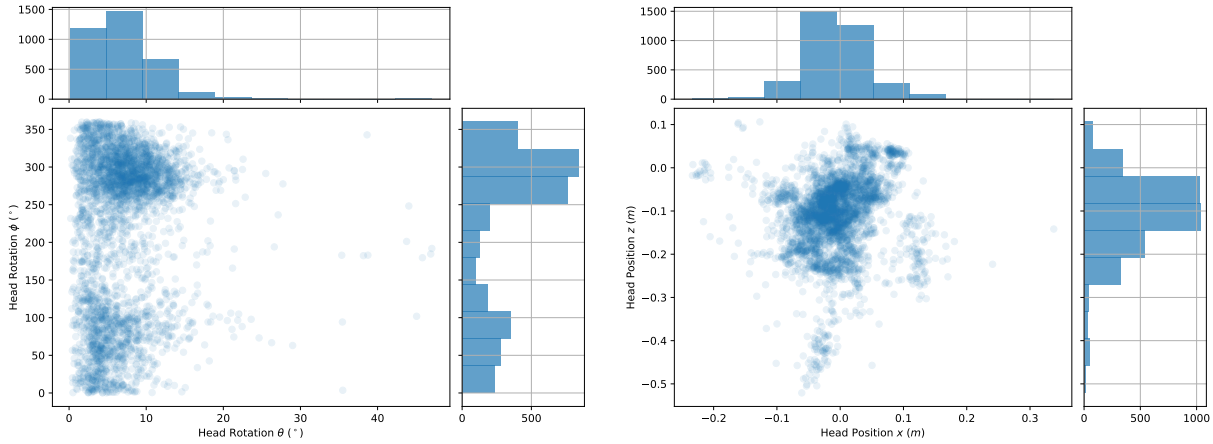


Figure 4.22: The final rotation and position on the horizontal plane for participant's head in each trial. Each point represents one trial, and darker areas have higher concentration of points.

of the initial position.

So far, we have analyzed the position and tilt of the containers and participants' heads in each trial. The remainder of this section examines the effect these parameters had on the success of the trials.

We included all trials with a defined correct answer (i.e. all conditions except for ($a1 = -0.3$, $a2 = 0.3$)), resulting in 3130 trials in total. We plotted the rotation and position variables for successful and unsuccessful trials individually and found that the distribution of the containers' final θ was different for successful and unsuccessful trials (figure 4.23). The other variables were similar for the successful and unsuccessful trials (the plots are included in appendix A.2.2). As the figure suggests, in the unsuccessful trials, the containers' tilts were closer to 0 or 90 degrees, while in the successful trials, the containers' tilt was closer to 45 degrees.

Lastly, we used logistic regression to evaluate the effect of the containers' final tilt on the success of the trials. As was suggested in figure 4.23, we used the absolute difference of the each container's tilt and 45 degrees as predictors and the success of each trial as the

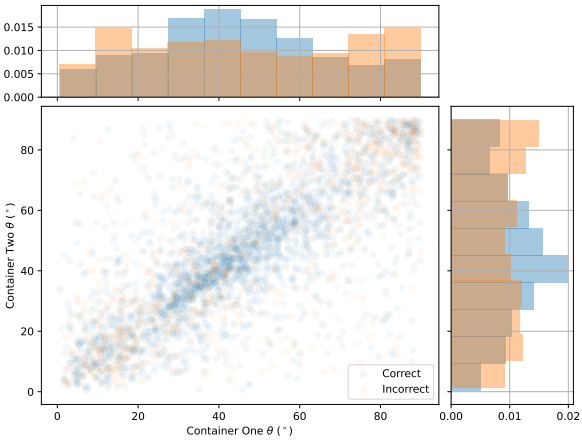


Figure 4.23: The final tilt of the two containers for successful (orange) and unsuccessful (blue) trials. Each point represents one trial, and darker areas have higher concentration of points. The distribution of containers' final tilt is different for successful and unsuccessful trials.

dependent variable. Also, we added the difference between the two containers' tilts as a predictor. Table 4.8 summarizes the regression results. As the results show, although the containers' tilts affected the trials' success, they could only explain 5 percent of the variance in the dependent variable.

4.5.4 AR-WLT Task for High and Low-scoring Participants

In this section, we compare the performance of low-scoring and high-scoring participants on AR-WLT. To do so, we only considered the AR-WLT-z score because it only used the trials where one simulation was correct, so choosing the right answer did not require imagining water's natural state and selecting the inaccurate simulation that was closer to it. Figure 4.24 shows the AR-WLT-z score distribution and five-point summary for high-scoring and low-scoring participants. As shown in the figure, the high-scoring participants scores were higher on average ($M(18) = 81.03$, $SD = 13.99$) than the low scoring participants ($M(17) = 55.39$, $SD = 9.34$). In fact, the low-scoring group's average performance was close to a random choice.

Table 4.8: The logistic regression results for analysing the relation between the containers' tilts and success of trials.

Dep. Variable:	correct	**No. Observations:**	3130
Model:	Logit	**Df Residuals:**	3126
Method:	MLE	**Df Model:**	3
converged:	True	**Pseudo R-squ.:**	0.05017
Covariance Type:	nonrobust	**Log-Likelihood:**	-1869.6
		LL-Null:	-1968.4
		LLR p-value:	1.448e-42

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	1.8146	0.092	19.777	0.000	1.635	1.994
abs(c1_theta - 45)	-0.0232	0.004	-6.344	0.000	-0.030	-0.016
abs(c2_theta - 45)	-0.0213	0.004	-5.805	0.000	-0.029	-0.014
abs(c2_theta - c1_theta)	-0.0148	0.003	-4.742	0.000	-0.021	-0.009

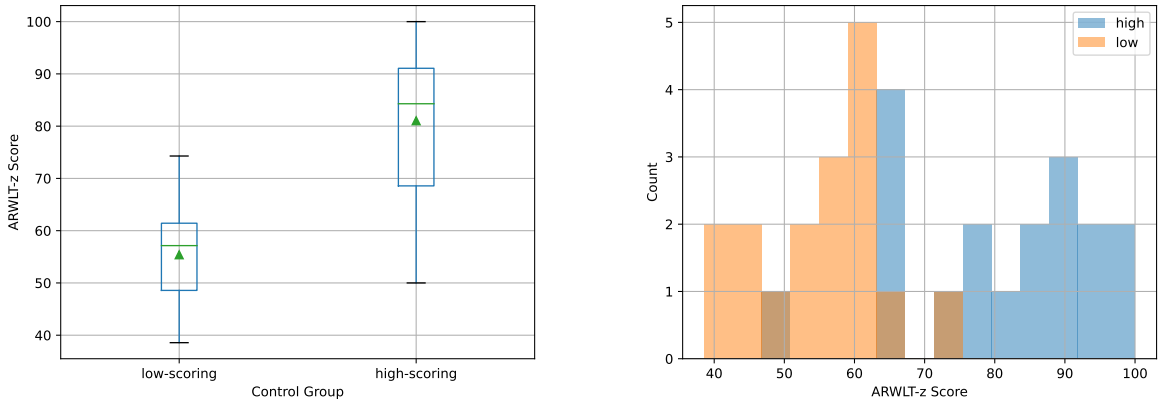


Figure 4.24: AR-WLT-z score five point summary and distribution for high scoring and low scoring participants.

The first step in formally comparing the two groups was to evaluate the two group's AR-WLT-z scores distribution normality. The Sharipo-Wilk test for normality failed to reject the null-hypothesis that the distributions were normal ($W = 0.946$, $p = .366$ for high-scoring and $W = 0.964$, $p = .707$ for low-scoring), and the quantile-quantile plots (figure 4.25) were close to linear, so we concluded the distributions were acceptably near normal. A

standard independent t-test was used to compare the two groups because the variances of the two groups were close (var(low-scoring)/$var_($high-scoring) = 2.24). The results showed the high-scoring group scored significantly higher than the low-scoring group ($t(33) = 6.340$, $p < 0.001$).

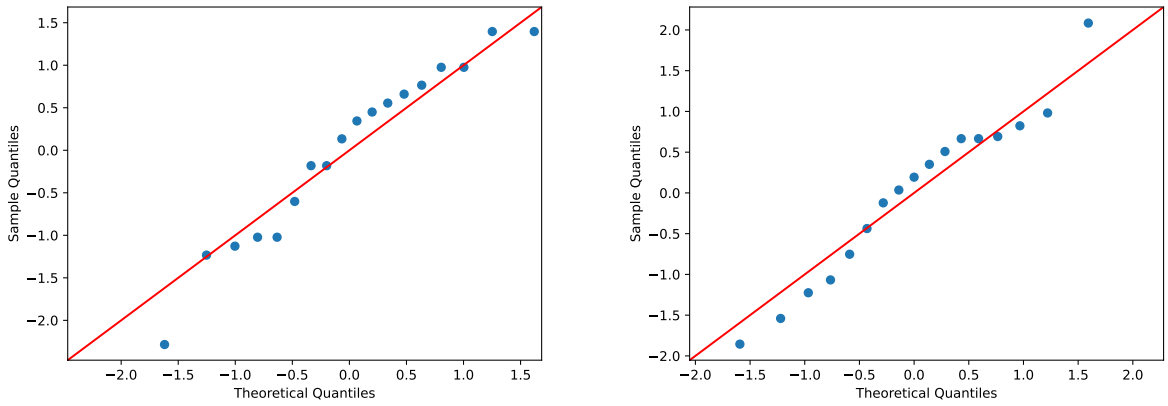


Figure 4.25: The AR-WLT-z score Q-Q plots for high-scoring (left) and low-scoring (right) participants.

Additionally, we performed linear regression with OLS error between the prescreening and AR-WLT scores. The results showed the prescreening score had a significant linear relationship with the AR-WLT-z score ($\beta = 1.737$, $p < 0.001$) and described 55% of its variance ($R^2 = 0.546$). Figure 4.26 shows the linear relationship between the AR-WLT-z and prescreening scores. The results of the linear regression are summarized in table 4.9.

The above results suggested individuals prone to the WLT error were less likely to find a liquid simulation that tilts with the container unrealistic. However, the performance on the WLT did not completely predict the performance on the AR-WLT. As discussed in section 4.5.1, the over-rotating simulations looked more realistic to an average participant. We will investigate next if it was the case for both low- and high-scoring groups.

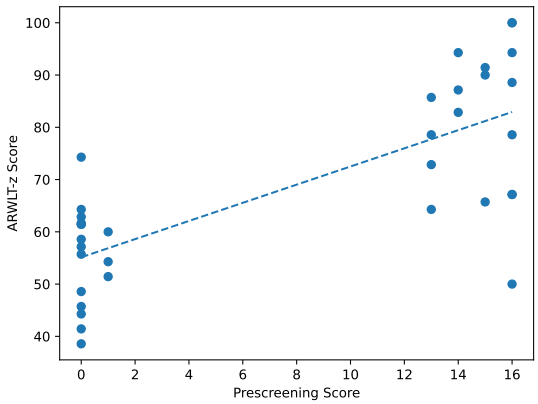


Figure 4.26: The AR-WLT-z score as a function of prescreening score.

Table 4.9: The linear regression results for analysing the relation between the prescreening score and the AR-WLT score.

Dep. Variable:	AR-WLT-z score	**R-squared:**	0.546
Model:	OLS	**Adj. R-squared:**	0.532
Method:	Least Squares	**F-statistic:**	39.66
No. Observations:	35	**Prob (F-statistic):**	4.03e-07
Df Residuals:	33	**Log-Likelihood:**	-135.61
Df Model:	1	**AIC:**	275.2
Covariance Type:	nonrobust	**BIC:**	278.3

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	55.1266	2.946	18.713	0.000	49.133	61.120
score_15	1.7373	0.276	6.297	0.000	1.176	2.299

The Effect of $a1$ and $a2$ on Low and High-scoring Participants

In this section we compare the high-scoring and low-scoring participants on different $a1$ and $a2$ settings. In table 4.10, average $a1$ rate (the number of times where simulation with $a = a1$ was chosen as the realistic simulation over total condition repetitions (10) as defined in section 4.5.1) for each condition is shown for high- and low-scoring participants.

We can see that for both groups, under-rotating simulations were easier to identify as

Table 4.10: Average $a1$ rate for different $a1$ and $a2$ combinations for high-scoring and low-scoring participants.

#	a1	a2	high-scoring	low-scoring
1	**0**	**1**	0.91	0.65
2	**0**	**0.7**	0.89	0.60
3	**0**	**0.5**	0.86	0.59
4	**0**	**0.3**	0.77	0.54
5	**0**	**-0.7**	0.82	0.52
6	**0**	**-0.5**	0.74	0.51
7	**0**	**-0.3**	0.68	0.46
8	**0.3**	**0.5**	0.77	0.59
9	**-0.3**	**-0.5**	0.76	0.48
10	**-0.3**	**0.3**	0.55	0.61

not realistic (comparing rows 2, 3, and 4 to 5, 6, and 7 respectively). This observation is especially interesting for the low-scoring group because their 2D drawings were more similar to the under-rotating settings. In fact, the low-scoring participants were most successful when comparing a correct simulation to a simulation that remained parallel to the bottom of the container (row 1), even though the erroneous simulation was more similar to their 2D depictions of the water orientation in the prescreening task. Rows 5, 6, and 7 of table 4.10 show that the low-scoring group did not find the over-rotating settings unrealistic, and their choices in the over-rotating versus correct settings were almost random. Figure 4.27 shows the average success rate in trials with one correct simulation and different $a2$s for high and low-scoring participants. We added the point (0, 0.5) to the dotted plots, reflecting expected chance performance.

4.5.5 AR-WLT and Control Task

As shown in section 4.4, the control task divided the participants into two groups whose scores were below 5° (referred to as the accurate group) and over 9° (referred to as the

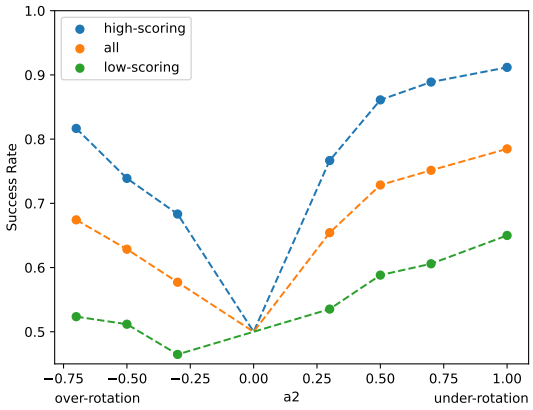


Figure 4.27: The success rate for settings with $a1 = 0$ and different $a2$s for low-scoring, high-scoring, and all participants.

inaccurate group). Similar to the previous section, we compared the AR-WLT-z scores of participants who scored below 5° (more accurate) and above 9° (less accurate) degrees. The one participant who aligned the surfaces vertically (discussed in sec 4.4) was excluded from this section's analyses. Figure 4.28 shows the AR-WLT-z distribution and five-point summary for the accurate and inaccurate groups.

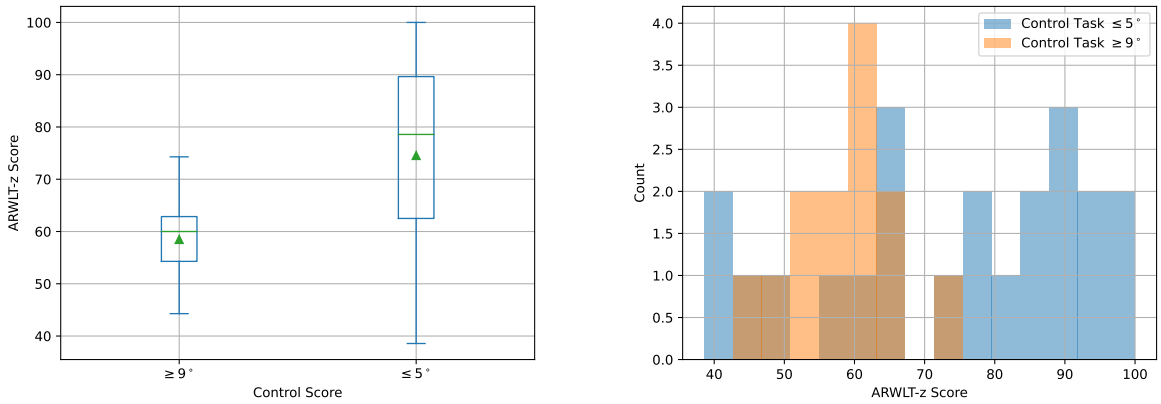


Figure 4.28: The AR-WLT-z score five point summary and distribution for participants separated by control task score.

As in the previous section, we tested the significance of the two groups' differences. The

Sharipo-Wilk normality test showed the two distributions were close to normal; the results were ($W = 0.934$, $p = 0.151$) for the accurate and ($W = 0.975$, $p = 0.945$) for the inaccurate group. The quartile-quartile plots are shown in figure 4.29.

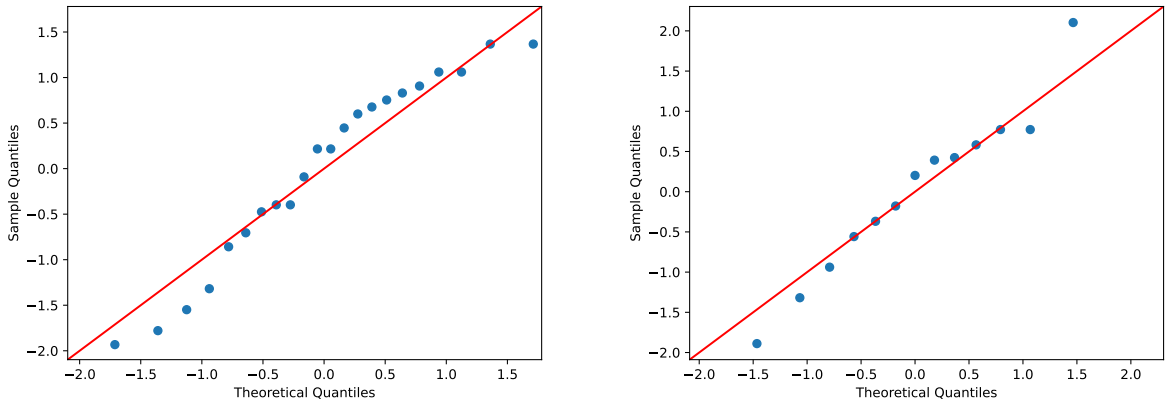


Figure 4.29: The AR-WLT-z score Q-Q plots for participants who were more accurate (left) and less accurate (right) on control task.

We used Welch's t-test because the variances were unequal ($\frac{var(\text{score}_{\geq 9})}{var(\text{score}_{\leq 5})} = 5.93$), which showed that the more accurate group's AR-WLT-z scores were significantly larger than the group who were less accurate on the control task ($t(32) = 3.489$, $p < 0.001$).

Finally, we used linear regression with ordinary least squares to evaluate the relation between the two task results, which showed a significant relationship between the two tasks ($\rho = -1.243$, $p = .007$). However, the control task result could only account for 21% of the variance in the AR-WLT task score. The results are shown in table 4.11 and the relationship between the two scores is shown in figure 4.30.

The Effect of $a1$ and $a2$ on Accurate and Inaccurate Participants

As discussed in section 4.4.1, a tilted container in the control task affected the participants' judgement of the horizontal surfaces, and on average, participants' answers were tilted towards the container's tilt, which is similar to the under-rotating settings. However,

Table 4.11: The linear regression results for analysing the relation between the control score and the AR-WLT score.

Dep. Variable:	AR-WLT-z score	**R-squared:**	0.209
Model:	OLS	**Adj. R-squared:**	0.185
Method:	Least Squares	**F-statistic:**	8.475
No. Observations:	34	**Prob (F-statistic):**	0.00651
Df Residuals:	32	**Log-Likelihood:**	-141.31
Df Model:	1	**AIC:**	286.6
Covariance Type:	nonrobust	**BIC:**	289.7

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	76.0186	3.840	19.798	0.000	68.198	83.840
control_score	-1.2437	0.427	-2.911	0.007	-2.114	-0.374

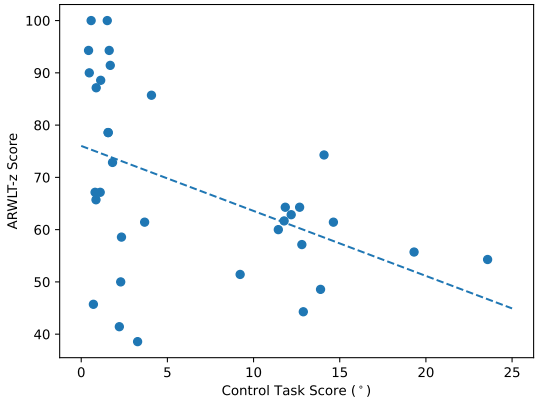


Figure 4.30: The AR-WLT-z score as a function of control task score.

section 4.5.1 concluded that under-rotating water simulations looked less realistic to participants on average. In this section, we analyzed the AR-WLT performance of the participants who were not accurate in the control task to see if they found the under-rotating settings (which they perceived as horizontal in the control task) more natural.

Figure 4.31 shows the five-point summary of the answers separately for accurate and inaccurate participants. As the figure suggests, the inaccurate participants' answers were tilted towards the container tilt. However, the variance of responses was high, especially in

the settings where the container tilt was twenty degrees.

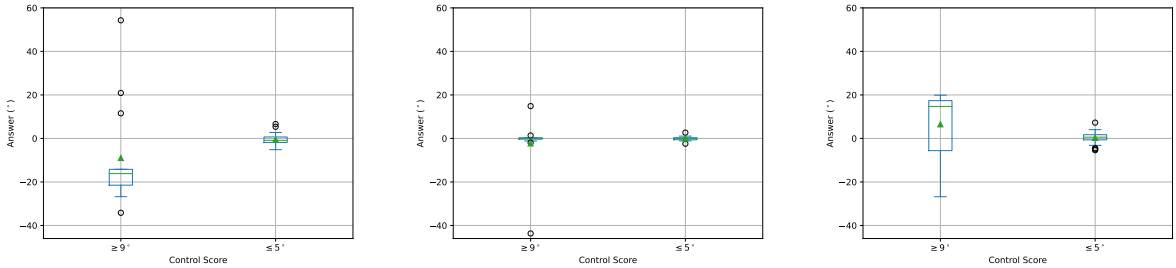


Figure 4.31: Control answer five point summary for more accurate and less accurate participants for container tilts -20, 0, and 20 from left to right.

We saw that the inaccurate participants were prone to perceiving surfaces tilted in the same direction as the container as horizontal. However, as shown in table 4.12, this orientation illusion did not cause them to favour the under-rotating settings over the horizontal or over-rotating settings. In fact, row 10 in the table shows that when choosing between the two over-rotating and under-rotating simulations with the same a value, the participants who were less accurate on the control task favoured over-rotating to the under-rotating simulations.

Table 4.12: Average $a1$ rate for different $a1$ and $a2$ combinations for participants whose control score was below 5° (accurate) and above 9° (inaccurate).

#	a1	a2	score $\leq 5°$	score $\geq 9°$
1	**0**	**1**	0.82	0.71
2	**0**	**0.7**	0.79	0.67
3	**0**	**0.5**	0.77	0.65
4	**0**	**0.3**	0.70	0.55
5	**0**	**-0.7**	0.74	0.55
6	**0**	**-0.5**	0.71	0.51
7	**0**	**-0.3**	0.65	0.45
8	**0.3**	**0.5**	0.70	0.66
9	**-0.3**	**-0.5**	0.70	0.48
10	**-0.3**	**0.3**	0.52	0.65

This observation contradicts the hypothesis that participants who did poorly on AR-WLT chose a tilted water surface as realistic because they perceived it as horizontal. We concluded that although there was a correlation between the control and AR-WLT tasks, the error on the AR-WLT task was not directly caused by participants' inability to evaluate the water surface's horizontality. If that were the case, the under-rotating settings would have looked more natural rather than over-rotating.

4.5.6 Discussion

We observed that in a dynamic setting (AR-WLT), it was easier for low-scoring participants to identify under-rotating anomalous simulations, although the under-rotating condition was similar to their abstract drawings. Moreover, although participants' performance in AR-WLT correlated with their control task performance, a misconception of the horizontal surface could not account for a low AR-WLT score because the over-rotating simulations were more likely to be mistakenly chosen as realistic. In contrast, participants' answers to the control task were under-rotating relative to the container. The interaction data analysis suggested that most participants devised the strategy to leave the containers rotated to the same degree and compare the water orientation in the containers. Finally, participants did not immediately identify anomalous simulations, which suggests that an analytical approach was required to perform successfully in the AR-WLT, and relying on intuition did not suffice.

4.6 Analysis of the Correlation Between the Three Tasks

In sections 4.5.4 and 4.5.5, we showed participants' results on prescreening and control tasks correlated with their performance on the AR-WLT. Moreover, in section 4.4.2 we found a significant correlation between the prescreening and control tasks. In this section, we examined the three tasks together.

Figure 4.32 shows the pairwise Pearson's correlation between the three tasks. One outlier participant (discussed in section 4.4) has been removed from the data for the control task.

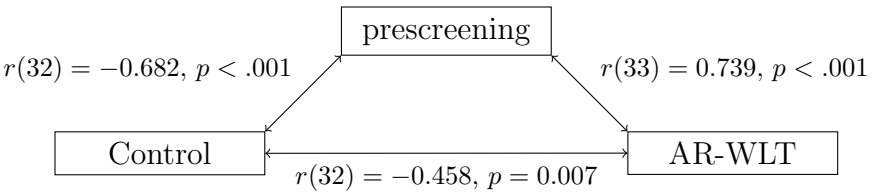


Figure 4.32: Pearson's correlation coefficient between the three tasks.

Regression analyses in section 4.5.4 and 4.5.5 showed that both prescreening and control tasks predicted the AR-WLT score. Tables 4.13 and 4.14 summarize the regression results.

Table 4.13: The regression summary of per-screening score(IV) and the AR-WLT score (DV).

R-squared:	0.546	**F-statistic:**		39.66		
Adj. R-squared:	0.532	**Prob (F-statistic):**		4.03e-07		
	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	55.1266	2.946	18.713	0.000	49.13	61.120
prescreening_score	1.7373	0.276	6.297	0.000	1.176	2.299

Table 4.14: The regression summary of control score(IV) and the AR-WLT score (DV).

R-squared:	0.209	**F-statistic:**		8.475		
Adj. R-squared:	0.185	**Prob (F-statistic):**		0.00651		
	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	76.0186	3.840	19.798	0.000	68.198	83.840
control_score	-1.2437	0.427	-2.911	0.007	-2.114	-0.374

We used multivariate regression analysis to investigate the simultaneous effect of the prescreening and control task on the AR-WLT score. More precisely, as the prescreening

and control tasks correlated significantly, we wanted to know if one task's results' effect on the AR-WLT could be explained by its correlation to the other predicting task.

When we simultaneously included both variables in the regression analysis, the control score's effect became insignificant ($\rho = 0.215$, $p = 0.638$). Moreover, as shown in table 4.13, the prescreening score could account for 55 percent of the variance in the AR-WLT score, which was the same in the multivariate model ($R^2 = 0.541$). The result of multivariate regression analysis with control and prescreening tasks as predictors of the AR-WLT task is summarized in table 4.15.

Table 4.15: The linear regression results for analysing the relation between the prescreening and control scores (IV) and the AR-WLT score (DV).

Dep. Variable:	scss_trials_z	**R-squared:**	0.541
Model:	OLS	**Adj. R-squared:**	0.511
Method:	Least Squares	**F-statistic:**	18.27
No. Observations:	34	**Prob (F-statistic):**	5.72e-06
Df Residuals:	31	**Log-Likelihood:**	-132.06
Df Model:	2	**AIC:**	270.1
Covariance Type:	nonrobust	**BIC:**	274.7

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	52.7983	5.736	9.205	0.000	41.100	64.497
prescreening_score	1.8523	0.391	4.733	0.000	1.054	2.650
control_score	0.2150	0.452	0.476	0.638	-0.707	1.137

The above results show the effect of the control task disappeared when prescreening task was added to the model (without the prescreening: table 4.14 ($\beta_1 = -1.24$), with the prescreening: table 4.15 ($\beta_1 = 0.21$)). This observation suggested that the prescreening results completely mediated between the control and AR-WLT results. We used bootstrapping mediation analysis to test the mediation significance with 1000 bootstrap iterations. The Pingouin python package[2] was used, which implements the Mediation package in R (Tingley

[2](`https://pingouin-stats.org/generated/pingouin.mediation_analysis.html`)

et al., 2014). As the Pingouin package is not widely used, we verified the results with R's Mediation package. Figure 4.33 shows the mediation analysis results. As shown in the figure, the direct effect of the control score on the AR-WLT score was not significant.

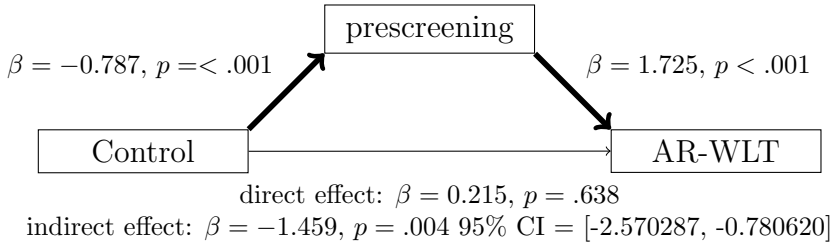


Figure 4.33: Analysis results for evaluating prescreening task as a mediator between control and AR-WLT results.

4.6.1 Discussion

Analyzing the relationship between all three tasks showed that prescreening scores could mediate the control score's correlation with the AR-WLT task. Although the relationship between the control and prescreening tasks is not necessarily causal, the mediation effect of the prescreening task was substantial. It showed that causes such as difficulty with the AR headset or environment did not affect the participants' performance significantly because if that were the case, there should have been a direct relationship between the two AR tasks.

Chapter 5

Discussion

5.1 Introduction

In this study, we recreated the Water Level Task in the Hololens2 AR environment to understand better why some adults do not draw horizontal water lines in a tilted container. We implemented a simple, fast liquid effect to examine humans' understanding of water orientation in a tilted container and compare their judgements in a 3D, immersive, interactive environment to 2D depictions.

We examined 120 undergraduate student participants on the traditional WLT and tested 35 high and low-scoring participants in the Augmented Reality environment. Revisiting the WLT cognitive task in a digital platform brought more detailed, accessible and analyzable data, which allowed us to understand the overall patterns and successful and unsuccessful strategies in understanding and drawing the water orientation in the WLT. This section summarizes our findings and provides future directions for the study.

5.2 Prescreening

We observed that the two performance measures previously used by researchers in the conventional WLT (the mean absolute tilt of answers (Vasta et al., 1993) and the number of successful trials (Liben, 1978)) highly correlated.

Similar to previous research (Vasta & Liben, 1996), gender correlated with participants' performance. Moreover, we found that the participants' devices' screen size also affected their performance. However, gender and screen size only accounted for 8% of variance in participants' scores, suggesting individual factors, such as knowledge of physics or one's approach to solving the puzzles, were of more impact and importance. Additionally, as previous research suggests (McAfee & Proffitt, 1991; Pascual-Leone & Morra, 1991; Vasta et al., 1993) we observed that shape and rotation of the containers affected the average participant's success rate or the puzzles' difficulty.

However, about a third of our participants drew lines parallel to the base of the container and did not change their strategy throughout the experiment. This observation contrasts Vasta et al. (1993)'s claim that most adults do not make such errors.

5.3 The Effect of the Realistic Environment

Studies have reported that individuals perform better in intuitive physics tasks when the task is more familiar and depicted more realistically (Battaglia et al., 2013; Kaiser et al., 1986; Schwartz & Black, 1999; Ullman et al., 2017). On the contrary, in our study, the participants' judgements were not more accurate in the more realistic environment (AR). We observed that participants' understanding of what was physically correct was tolerant to minor inaccuracies even if the underlying physics was known to them (high-scoring participants). Modern physics simulation tools take advantage of this tolerance to simplify complex

physics equations for faster simulations.

Kaiser et al. (1992) proposed that animation can improve one's performance in intuitive tasks if the task is easy enough (only one relative parameter needs to be accounted for) and the animation draws attention to the salient parameter. It could be argued that our dynamic setting added irrelevant parameters, such as water surface movement, so it did not draw participants' attention to gravity and the orientation of water.

5.4 Individual Inconsistency

Researchers have reported that subjects' performance in similar tasks varies greatly in abstract and more realistic settings (Cook & Breedin, 1994; Schwartz & Black, 1999). Kaiser et al. (1985) observed that participants did not choose anomalous simulations similar to their 2D depictions of trajectories of moving objects. Our data also showed that although participants' performance in the AR setting correlated with their performance in the abstract setting (prescreening), their choice of realistic water orientation was inconsistent in the abstract and AR settings. More specifically, low-scoring participants preferred over-rotating AR simulations to under-rotating simulations even though their drawn lines were consistent with the most exaggerated under-rotating setting.

5.4.1 Over-rotating and Under-rotating

Previous findings have been inconsistent on the tilt of the drawn lines with respect to the container's tilt. McAfee & Proffitt (1991) reported that the drawn lines were tilted towards the container's tilt, and Vasta & Liben (1996) reported adults, especially individuals who imagined water in a dynamic state, drew lines tilted in the opposite direction of the container's tilt (i.e. towards the edge of the container) (from Vasta 1994). We found that the high-scoring participants' answers (scoring 13 and above) were tilted towards the container

tilt. This suggests that although they drew horizontal lines (within 15° of horizontal) in at least 13 out of 16 puzzles, their answers were likely affected by the Tilt Illusion. However, low-scoring participants' errors were more likely caused by using a wrong heuristic rather than the Tilt Illusion effect, as their answers were within 15° of the base of the container. For mid-scoring participants (scoring 2 to 12 out of 16), the answers were not significantly more likely to be tilted in either direction.

Howard (1978) reported that participants who did not draw the correct water line found videos in which the water surface over-rotated to 20° natural. However, the under-rotating water looked natural for up to 10°, suggesting participants had more tolerance to over-rotating irregularity. In contrast, McAfee & Proffitt (1991) reported that over-rotating settings were easiest for participants to identify in anomalous pictures.

We found that in the AR setting, both low and high-scoring participants preferred the over-rotating simulations to the under-rotating ones. However, in the prescreening task, both groups' answers were, on average, tilted in the same direction as the container tilt. Our finding, along with the previous research finding, suggests that, in more realistic settings, over-rotating simulations are preferred to under-rotating simulations, and in the abstract line drawing settings, under-rotating water lines are preferred to over-rotating ones.

The difference in preference for over-rotating simulations in more realistic settings and under-rotating water lines in abstract line drawing settings may be due to differences in perceptual processing and cognitive mechanisms involved in these tasks. In more realistic settings, participants may rely more on visual and depth cues and feedback from the environment to make judgments, while in abstract line drawing settings, participants may rely more on cognitive heuristics and prior knowledge to interpret the task.

5.5 Learning

Vasta et al. (1996) reported that participants' performance improved in the second half of the conventional WLT. We also observed performance improvement during the 16 conventional tasks. Although, 55% of participants consistently drew lines parallel to the base of the container or the horizontal (suggesting their strategy did not change during the task). We did not find reports of performance improvement in abstract tasks in the absence of feedback in other intuitive physics tasks.

We did not find evidence of a significant performance improvement due to learning in the AR-WLT. However, the average time spent on the task reduced quickly during the first few trials, suggesting participants got more proficient in performing the task procedure.

5.6 Intuitive Physics Models

As discussed in section 1.3.3, two (not mutually exclusive) models of human intuitive understanding of physics have been introduced. Some researchers proposed that humans rely on piecemeal heuristics to reason about physical events (Cohen, 2006; Davis & Marcus, 2016; Proffitt & Gilden, 1989; Proffitt & Kaiser, 2006; Todd & Warren Jr, 1982). Other researchers suggest a probabilistic, simulation/prediction based model for human reasoning about everyday physics (Bates et al., 2015; Battaglia et al., 2013; Firestone & Scholl, 2017; Hegarty, 2004; J. Kubricht et al., 2016; Smith et al., 2013).

Our observations suggest that participants used different strategies in the AR-WLT and prescreening tasks.

The prescreening data suggested that in the conventional WLT, low and high-scoring participants relied on simple heuristics to draw the lines. As 55% of participants consistently drew lines parallel to the base of the container or horizontally. Moreover, performance

improvement during the task suggests that mid-scoring participants used analysis to some extent to find and use the correct heuristic.

We observed that in the AR-WLT, the gap between high and low-scoring participants' performance was not as definite as it was in the pre-screening task. This observation suggests that participants' knowledge of the horizontality principle did not directly affect their intuition of water orientation. Previous research has also shown that although knowledge of a physics principle highly correlates with one's performance on the task, it does not immediately affect one's intuitive comprehension of physical events (Proffitt et al., 1990). Moreover, we expected the participants to recognize the wrong simulations, especially in exaggerated cases immediately. The fact that the simulations did not instantly feel unnatural suggested that participants' intuitive understanding of the water surface was tolerant to minor errors.

Finally, participants did not choose the AR simulations similar to their 2D drawings indicating that even for high-scoring participants, what looked natural was different in the 2D and AR settings. Moreover, the low-scoring participants' inconsistency between conventional and AR WLT suggests that different cognitive and perceptual factors affected their errors in the two tasks.

5.7 Type of Anomaly

Among the many physical principles that form our expectation of how an object reacts to our actions, some are more crucial than others for our feeling of natural interaction. In the water-level case, we believe the amount of the water plane's tilt relative to the container's tilt is not as crucial as other factors, such as the orientation of the water tilt. In other words, our wrong simulations did not immediately feel unnatural because when the container tilted, the water touched the correct, expected point on the container's edge (i.e. the point toward which the container was tilted). The fact that water pours out of the edge toward which the

container tilts is more crucial than the tilt degree at which the water touches the edge (and pours out) because, in our daily interactions with the container, we do not need to predict the latter. We tilt the container gradually until the water touches the edge of the container and pours out. This hypothesis explains why incorrect AR simulations did not immediately feel unnatural; however, it cannot explain the 2D drawings of low-scoring participants.

5.8 Recommendations for Future Work

In section 4.5.4, we saw that the low-scoring participants did not find the simulations that were similar to their drawings more realistic than a correct simulation. Repeating a 2D WLT after the participants interact with the AR simulations would be interesting to see if their answers would change.

As the data showed, slight physics alterations did not feel unrealistic. However, we did not investigate if they would make participants' predictions more difficult. A future study can ask participants to tilt containers till the water is just about to pour out and measure their speed and accuracy on the correct and wrong simulations. Moreover, it can be investigated if the participants would adjust their behaviour or expectations when a wrong simulation repeats multiple times. Additionally, it would be interesting to know whether having to tilt the container till the liquid is about to pour out makes it easier or harder for participants to judge if the simulation was correct.

In our experiment, the wrong physics simulation did not immediately feel wrong to participants. We believe it was because, in our simulations, the liquid would pour out of the expected point on the edge of the container. A possible follow-up task is to include simulations where the water stays close to parallel to the bottom of the container in all tilt degrees (similar to low-scoring participants' drawings) and never pours out. Alternatively, in a more complex case, the same experiment can be repeated with simulations where the water

plane's normal does not tilt toward the container's y axis, resulting in the liquid touching a wrong point on the edge of the container. We believe that alteration would feel more unnatural.

Finally, in section 5.7, we argued that some physical rules are of more importance whose alterations would result in a more unnatural feeling. Other physics tasks could be used to test what rules are more crucial than the others; for example, a task in which the participants are asked to catch a ball, where the physics of the ball's movement is incorrect.

Appendix A

A.1 Prescreening Data Analysis

Additional prescreening data analyses are included in this section.

A.1.1 Comparing Performance Measures

As discussed in section 1.2, researchers have used either the number of successful trials or the average absolute tilt from the horizontal over the trials to measure performance on the conventional WLT. We used the number of successful trials with a threshold 15° to measure participants' performance and identify high and low-scoring participants (section 4.2). We observed that the number of successful trials with a threshold 15° highly correlated with the average absolute tilt from horizontal ($R^2 = 0.94$). This observation suggests that the choice of the performance measure did not significantly affect our analyses. Table A.1 summarizes linear regression between the number of successful trials and the average absolute tilt from horizontal for our participants. The average absolute value of the answer tilt for high-scoring participants was ($M(38) = 6.04°$, $sd = 3.95$) with a maximum of 16.16° and a minimum of 1.23°, and for low-scoring participants, it was ($M(42) = 48.91°$, $sd = 3.36$) with maximum and minimum of 36.01° and 56.87°, respectively.

Figure A.1 shows the number of successful trials plotted against the average absolute tilt

Table A.1: The linear regression results between the average absolute tilt and number of successful trials in the prescreening task for each participant.

Dep. Variable:	abs_answer_tilt	**R-squared:**	0.944
Model:	OLS	**Adj. R-squared:**	0.944
Method:	Least Squares	**F-statistic:**	1969.
No. Observations:	118	**Prob (F-statistic):**	1.28e-74
Df Residuals:	116	**Log-Likelihood:**	-341.80
Df Model:	1	**AIC:**	687.6
Covariance Type:	nonrobust	**BIC:**	693.1

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	48.8086	0.616	79.210	0.000	47.588	50.029
score_15	-2.8846	0.065	-44.377	0.000	-3.013	-2.756

for thresholds 5°, 10° and 15°. The linear regression line is also plotted for successful trials with a threshold of 15°.

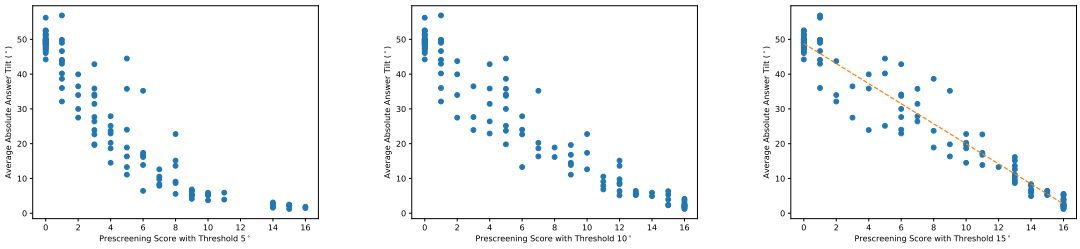


Figure A.1: Number of successful trials with thresholds 5°, 10° and 15° (from left to right) plotted against average absolute tilt. The linear regression line is plotted in the rightmost figure.

A.1.2 Low-scoring Participants Whose Answers Were not Parallel to the Bottom of the Container

Figure A.2 shows answers of low-scoring participants whose answers were not within 15° of the container tilt. The left-most participant appeared more accurate in the first four trials

and changed their strategy using inaccurate heuristics. The middle participant drew lines that were close to parallel to the sides of the container, the same as five of the right-most participant's answers. The right-most participant has drawn vertical lines in four trials, and their strategy is unclear in the rest of the trials.

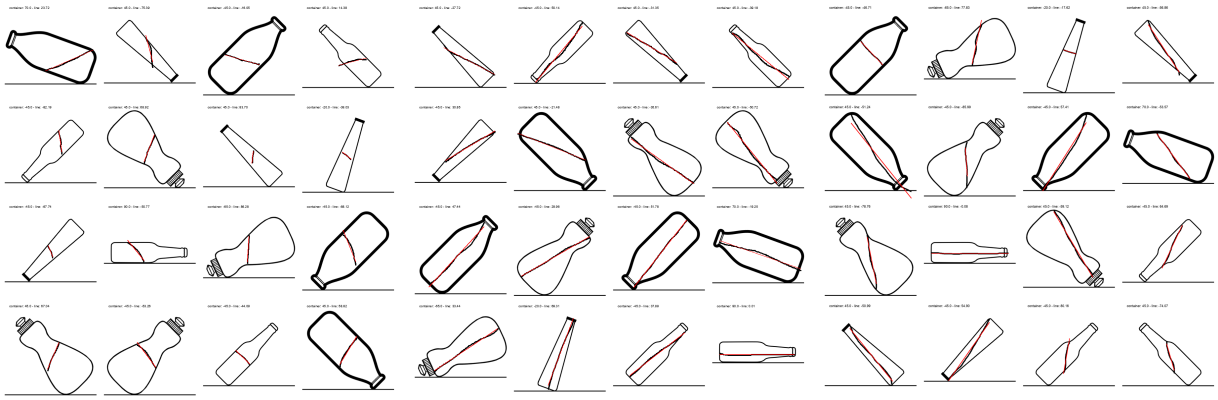


Figure A.2: The answers of low-scoring participants who did not draw lines parallel to the bottom of the container.

A.1.3 Further Tilt Illusion Analysis

Section 4.2.2 showed that tilt illusion could not account for low-scoring participants' errors. So, to further analyze the tilt illusion, we excluded low-scoring participants' data and evaluated the mid-scoring (the participants who scored between 2 and 13) and high-scoring participants' answers.

For each individual, we calculated the ratio of their answers tilted in the same direction as the container. On average, 0.57 ($SD = 0.27$) of mid-scoring and 0.61 ($SD = 0.17$) of high-scoring participants' answers were tilted in the same direction as the container. Among all mid-scoring and high-scoring participants ($N = 76$), the majority of 53 participants' answers were tilted towards the container tilt. Most answers were tilted opposite the container for the remaining 23 individuals (8 high-scoring and 15 mid-scoring). Related samples t-test showed high-scoring participants were significantly more likely to draw lines in the same direction

as container tilt ($t(37) = 3.93$, $p < .001$). However, for the mid-scoring participants, the ratio of answers tilted towards the container tilt was not significantly higher than the rest ($t(37) = 1.59, p = 0.060$). We concluded that high-scoring participants' answers were affected by the tilt illusion. However, the data was insufficient to show the same for mid-scoring participants' answers suggesting wrong heuristics could also have affected their answers.

A.1.4 The Success Rate for Each Puzzle

Table A.2 includes the prescreening puzzles sorted by the success rate of all participants (number of successful trials over the total number of participants). Comparing the eight puzzles with lower and highest success rates (the half top and bottom of the table, respectively) suggests the upright containers were easier for participants than the containers that were pointing downwards (only one upright container appears in the top half and one downward container appears in the bottom half).

A.1.5 The Absolute Average Answer Tilt for Each Puzzle

Figure A.3 shows the distribution of participants' answers' tilts from horizontal for each puzzle. In each plot, the orange line shows the horizontal line's tilt (answer tilt = 0°), and the red line(s) shows the tilt of a line parallel to the bottom of the container. As shown in the figure, there are two peaks around the orange and red lines for all puzzles, indicating that most answers were either approximately horizontal or parallel to the bottom of the container. However, some responses were neither tilted with the container nor horizontal.

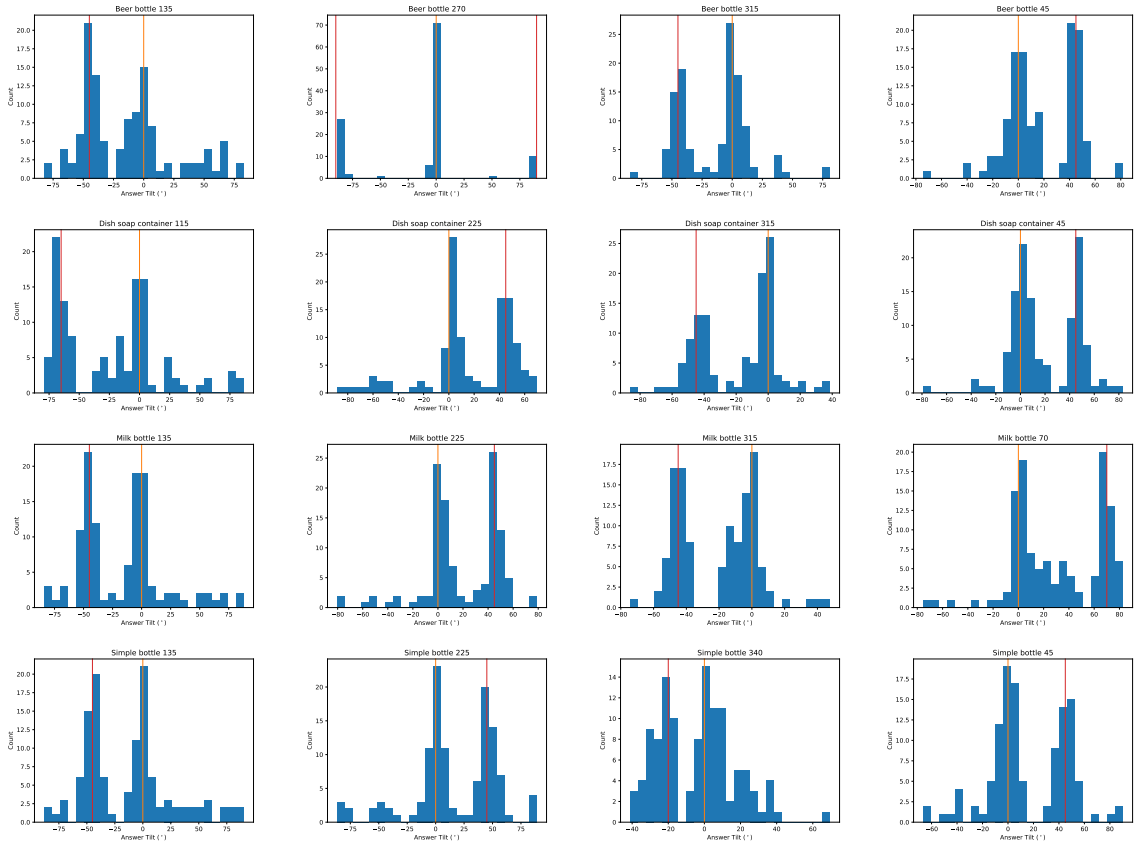


Figure A.3: The distribution of the participants' answers' tilts from horizontal for each puzzle. The orange and red lines indicate a horizontal line and a line parallel to the bottom of the container, respectively.

A.2 AR-WLT Data Analysis

A.2.1 AR-WLT Individual Successful Trials Binomial Test

Table A.3 is the result of binomial test for individual participants in the AR-WLT experiment.

A.2.2 AR-WLT Interaction Analysis Plots

The following plots show the final position and rotation of the two containers and head were not significantly different for successful and unsuccessful trials (Section 4.5.3).

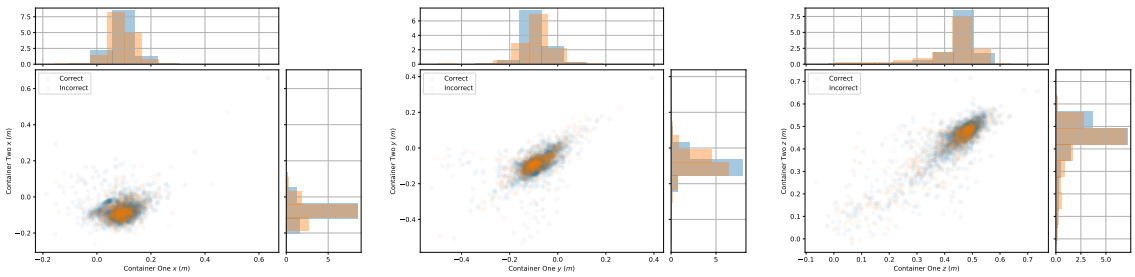


Figure A.4: The final position of the two containers for successful and unsuccessful trials. The histograms are normalized to sum to one.

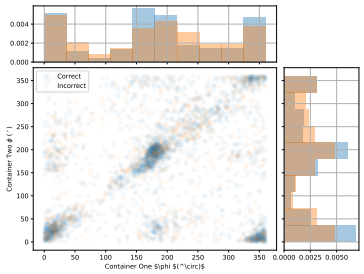


Figure A.5: The final tilt direction (ϕ) of the two containers for successful and unsuccessful trials. The histograms are normalized to sum to one.

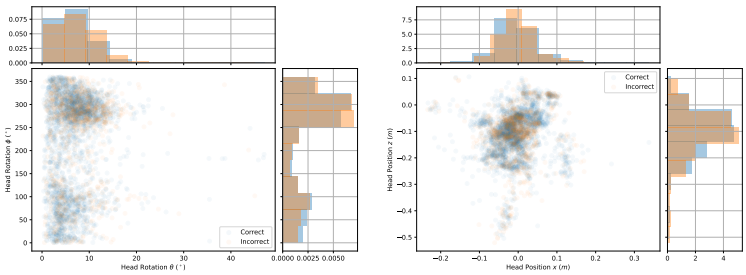


Figure A.6: The final head rotation (left) and position on the x and z axes (right) for the successful and unsuccessful trials. The histograms are normalized to sum to one.

Image	Container	Rotation	Success Rate
	Beer-bottle	135°	0.339
	Dish-soap-container	115°	0.339
	Simple-bottle	135°	0.356
	Dish-soap-container	225°	0.390
	Milk-Bottle	135°	0.398
	Milk-Bottle	70°	0.398
	Simple-bottle	340°	0.415
	Simple-bottle	225°	0.424
	Milk-Bottle	225°	0.432
	Beer-bottle	45°	0.458
	Milk-bottle	315°	0.475
	Simple-bottle	45°	0.492
	Beer-bottle	315°	0.517
	Dish-soap-container	315°	0.517
	Dish-soap-container	45°	0.517
	Beer-Bottle	270°	0.653

Table A.2: The success rate of all participants for different prescreening puzzles.

	#Success	Total	Proportion Estimate	Two sided 95% CI
1	77	90	0.86	[0.77, 0.92]
2	67	90	0.74	[0.64, 0.83]
3	54	90	0.60	[0.49, 0.70]
4	81	90	0.90	[0.82, 0.95]
5	59	90	0.66	[0.55, 0.75]
6	44	90	0.49	[0.38, 0.60]
7	60	90	0.67	[0.56, 0.76]
8	52	90	0.58	[0.47, 0.68]
9	39	90	0.43	[0.33, 0.54]
10	50	90	0.56	[0.45, 0.66]
11	80	80	1.00	[0.95, 1.00]
12	34	90	0.38	[0.28, 0.49]
13	80	90	0.89	[0.81, 0.95]
14	59	90	0.66	[0.55, 0.75]
15	67	90	0.74	[0.64, 0.83]
16	46	90	0.51	[0.40, 0.62]
17	81	90	0.90	[0.82, 0.95]
18	50	90	0.56	[0.45, 0.66]
19	70	90	0.78	[0.68, 0.86]
20	86	90	0.96	[0.89, 0.99]
21	78	90	0.87	[0.78, 0.93]
22	46	90	0.51	[0.40, 0.62]
23	47	80	0.59	[0.47, 0.70]
24	80	90	0.89	[0.81, 0.95]
25	54	90	0.60	[0.49, 0.70]
26	54	90	0.60	[0.49, 0.70]
27	73	90	0.81	[0.71, 0.89]
28	71	90	0.79	[0.69, 0.87]
29	46	90	0.51	[0.40, 0.62]
30	40	90	0.44	[0.34, 0.55]
31	86	90	0.96	[0.89, 0.99]
32	55	90	0.61	[0.50, 0.71]
33	51	90	0.57	[0.46, 0.67]
34	45	90	0.50	[0.39, 0.61]
35	58	90	0.64	[0.54, 0.74]

Table A.3: Binomial test confidence interval for each participant in the AR-WLT.

References

Alex, M., & Fischer, J. (2020). When it all falls down: the relationship between intuitive physics and spatial cognition. *Cognitive Research*, *5*(1).

Annett, M. (1994). Handedness as a continuous variable with dextral shift: sex, generation, and family handedness in subgroups of left-and right-handers. *Behavior genetics*, *24*(1), 51–63.

Baenninger, M., & Newcombe, N. (1995). Environmental input to the development of sex-related differences in spatial and mathematical ability. *Learning and Individual Differences*, *7*(4), 363–379.

Baillargeon, R. (1996). Infants' understanding of the physical world. *Journal of the Neurological Sciences*, *143*(1-2), 199–199.

Barhorst-Cates, E. M., Creem-Regehr, S. H., Stefanucci, J. K., Gardner, J., Saccomano, T., & Wright, C. (2020). Spatial reference frame but neither age nor gender predict performance on a water-level task in 8-to 11-year-old children. *Perception*, *49*(11), 1200–1212.

Barsky, R. D., & Lachman, M. E. (1986). Understanding of horizontality in college women: Effects of two training procedures. *International Journal of Behavioral Development*, *9*(1), 31–43.

Bates, C. J., Battaglia, P. W., Yildirim, I., & Tenenbaum, J. B. (2015). Humans predict liquid dynamics using probabilistic simulation. In *Cogsci*.

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. (2019). Modeling human intuitions about liquid flow with particle-based simulation. *PLoS computational biology*, *15*(7), e1007210.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Bridges, D., Pitiot, A., MacAskill, M., & Peirce, J. (2020, 07). The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. doi: doi: 10.7717/peerj.9414

Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects. *Cognition*, *9*(2), 117–123.

Casey, M. B. (1996). Understanding individual differences in spatial ability within females: A nature/nurture interactionist framework. *Developmental Review*, *16*(3), 241–260.

Cohen, A. L. (2006). Contributions of invariants, heuristics, and exemplars to the visual perception of relative mass. *Journal of experimental psychology: human perception and performance*, *32*(3), 574.

Cohen, A. L., & Ross, M. G. (2009). Exploring mass perception with markov chain monte carlo. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1833.

Collaer, M. L., & Hines, M. (1995). Human behavioral sex differences: A role for gonadal hormones during early development? *Psychological bulletin*, *118*(1), 55.

Cook, N. J., & Breedin, S. D. (1994). Constructing naive theories of motion on the fly. *Memory & Cognition*, *22*(4), 474–493.

Corbett, J. E., & Enns, J. T. (2006). Observer pitch and roll influence: the rod and frame illusion. *Psychonomic bulletin & review*, *13*(1), 160–165.

Coren, S., & Hoy, V. S. (1986). An orientation illusion analog to the rod and frame: Relational effects in the magnitude of the distortion. *Perception & Psychophysics*, *39*(3), 159–163.

Davis, E., & Marcus, G. (2016). The scope and limits of simulation in automated reasoning. *Artificial Intelligence*, *233*, 60–72.

De Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*(1), 1–12.

Ekstrom, R. B., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests, 1976*. Educational testing service.

Firestone, C., & Scholl, B. (2017). Seeing physics in the blink of an eye. *Journal of Vision*, *17*(10), 203–203.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the national academy of sciences*, *113*(34), E5072–E5081.

Goodenough, D. R., Oltman, P. K., Sigman, E., Rosso, J., & Mertz, H. (1979). Orientation contrast effects in the rod-and-frame test. *Perception & Psychophysics*, *25*(5), 419–424.

Hecht, H., & Proffitt, D. R. (1995). The price of expertise: Effects of experience on the water-level task. *Psychological Science*, *6*(2), 90–95.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in cognitive sciences*, *8*(6), 280–285.

Hegarty, M., & Sims, V. K. (1994). Individual differences in mental animation during mechanical reasoning. *Memory & Cognition*, *22*(4), 411–430.

Hespos, S. J., Ferry, A. L., Anderson, E. M., Hollenbeck, E. N., & Rips, L. J. (2016). Five-month-old infants have general knowledge of how nonsolid substances behave and interact. *Psychological Science*, *27*(2), 244–256.

Howard, I. P. (1978). Recognition and knowledge of the water-level principle. *Perception*, *7*(2), 151–160.

Jiang, C. (2015). *The material point method for the physics-based simulation of solids and fluids*. University of California, Los Angeles.

Kaiser, M. K., Jonides, J., & Alexander, J. (1986). Intuitive reasoning about abstract and familiar physics problems. *Memory & Cognition*, *14*(4), 308–312.

Kaiser, M. K., Proffitt, D. R., & Anderson, K. (1985). Judgments of natural and anomalous trajectories in the presence and absence of motion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 795.

Kaiser, M. K., Proffitt, D. R., Whelan, S. M., & Hecht, H. (1992). Influence of animation on dynamical judgments. *Journal of experimental Psychology: Human Perception and performance*, *18*(3), 669.

Kalichman, S. C. (1988). Individual differences in water-level task performance: A component-skills analysis. *Developmental Review*, *8*(3), 273–295.

Kawabe, T., Maruya, K., Fleming, R. W., & Nishida, S. (2015). Seeing liquids from visual motion. *Vision research*, *109*, 125–138.

Kenyon, J. (1984). Paper-and-pencil tests of piaget's water-level test: Sex differences and test modality. *Perceptual and motor skills*, *59*(3), 739–742.

Kestenbaum, R., Termine, N., & Spelke, E. S. (1987). Perception of objects and object boundaries by 3-month-old infants. *British journal of developmental psychology*, *5*(4), 367–383.

Krist, H., Fieberg, E. L., & Wilkening, F. (1993). Intuitive physics in action and judgment: The development of knowledge about projectile motion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 952.

Kubricht, J., Jiang, C., Zhu, Y., Zhu, S.-C., Terzopoulos, D., & Lu, H. (2016). Probabilistic simulation predicts human performance on viscous fluid-pouring problem. In *Cogsci.*

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, *21*(10), 749–759.

Liben, L. S. (1978). Performance on piagetian spatial tasks as a function of sex, field dependence, and training. *Merrill-Palmer Quarterly of Behavior and Development*, *24*(2), 97–110.

Liben, L. S., & Golbeck, S. L. (1984). Performance on piagetian horizontality and verticality tasks: Sex-related differences in knowledge of relevant physical phenomena. *Developmental Psychology*, *20*(4), 595.

Loh, M. N., Kirsch, L., Rothwell, J. C., Lemon, R. N., & Davare, M. (2010). Information about the weight of grasped objects from vision and internal models interacts within the primary motor cortex. *Journal of Neuroscience*, *30*(20), 6984–6990.

McAfee, E. A., & Proffitt, D. R. (1991). Understanding the surface orientation of liquids. *Cognitive Psychology*, *23*(3), 483–514.

McCloskey, M. (1983). Intuitive physics. *Scientific american*, *248*(4), 122–131.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, *210*(4474), 1139–1141.

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: the straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 636.

Michotte, A. (2017). *The perception of causality.* Routledge. (Originally published 1963)

Microsoft. (2022). *Hololens2.* [Online]. Available from: `https://www.microsoft.com/en-us/hololens`. (Last checked on Aug 01, 2022)

Mircosoft. (2022). *Mixed reality toolkit.* [Online]. Available from: `https://docs.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk2/?view=mrtkunity-2022-05`. (Last checked on Jun 01, 2022)

Monaghan, J. J. (2005). Smoothed particle hydrodynamics. *Reports on progress in physics*, *68*(8), 1703.

Myer, K. A., & Hensley, J. H. (1984). Cognitive style, gender, and self-report of principle as predictors of adult performance on piaget's water level task. *The Journal of genetic psychology*, *144*(2), 179–183.

Oakes, L. M. (1994). Development of infants' use of continuity cues in their perception of causality. *Developmental Psychology*, *30*(6), 869.

Pascual-Leone, J., & Morra, S. (1991). Horizontality of water level: A neo-piagetian developmental review. *Advances in child development and behavior*, *23*, 231–276.

Piaget, J., & Inhelder, B. (1956). *Child's conception of space* (F. Langdon & J. Lunzer, Trans.). London: Routledge & Kegan Paul. (Original work published 1948)

Proffitt, D. R., & Gilden, D. L. (1989). Understanding natural dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 384.

Proffitt, D. R., & Kaiser, M. K. (2006). Intuitive physics. *Encyclopedia of cognitive science*.

Proffitt, D. R., Kaiser, M. K., & Whelan, S. M. (1990). Understanding wheel dynamics. *Cognitive psychology*, *22*(3), 342–373.

PTC Inc. (2022). *Vuforia hololens2 sample (version ...) [unity asset].* [Online]. Available from: `https://assetstore.unity.com/packages/templates/packs/vuforia-hololens-2-sample-101553#description`. (Last checked on Jun 01, 2022)

Quaiser-Pohl, C., Lehmann, W., & Eid, M. (2004). The relationship between spatial abilities and representations of large-scale space in children—a structural equation modeling analysis. *Personality and Individual Differences*, *36*(1), 95–107.

Ranney, M., & Thagard, P. (1988). *Explanatory coherence and belief revision in naive physics* (Tech. Rep.). PITTSBURGH UNIV PA LEARNING RESEARCH AND DEVELOPMENT CENTER.

Rebelsky, F. (1964). Adult perception of the horizontal. *Perceptual and Motor Skills*, *19*(2), 371–374.

Robert, M., & Harel, F. (1996). The gender difference in orienting liquid surfaces and plumblines: Its robustness, its correlates, and the associated knowledge of simple physics. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *50*(3), 280.

Robert, M., & Morin, P. (1993). Gender differences in horizontality and verticality representation in relation to initial position of the stimuli. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *47*(3), 507.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2009). A bayesian framework for modeling intuitive dynamics. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 1–6).

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, *120*(2), 411.

Schwartz, D. L. (1999). Physical imagery: Kinematic versus dynamic models. *Cognitive Psychology*, *38*(3), 433-464. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0010028598907022` doi: doi: https://doi.org/10.1006/cogp.1998.0702

Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(1), 116.

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th python in science conference.*

SHAPIRO, S. S., & WILK, M. B. (1965, dec). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611. Retrieved from `https://doi.org/10.1093/biomet/52.3-4.591` doi: doi: 10.1093/biomet/52.3-4.591

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Sholl, M. J. (1989). The relation between horizontality and rod-and-frame and vestibular navigational performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(1), 110.

Sholl, M. J., & Liben, L. S. (1995). Illusory tilt and euclidean schemes as factors in performance on the water-level task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(6), 1624.

Signorella, M. L., & Jamison, W. (1978). Sex differences in the correlations among field dependence, spatial ability, sex role orientation, and performance on piaget's water-level task. *Developmental Psychology*, *14*(6), 689.

Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems. *In R. S. Siegler (Ed.), Children's thinking: What develops? (pp. 325–348).*

Smith, K. A., Battaglia, P., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

Spelke, E. S., Kestenbaum, R., Simons, D. J., & Wein, D. (1995). Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British journal of developmental psychology*, *13*(2), 113–142.

Thomas, H., & Jamison, W. (1975). On the acquisition of understanding that still water is horizontal. *Merrill-Palmer Quarterly of Behavior and Development*, *21*(1), 31–44.

Thomas, H., & Jamison, W. (1981). A test of the x-linked genetic hypothesis for sex differences on piaget's water-level task. *Developmental Review*, *1*(3), 274–283.

Thomas, H., Jamison, W., & Hummel, D. D. (1973). Observation is insufficient for discovering that the surface of still water is invariantly horizontal. *Science*, *181*(4095), 173–174.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). *Mediation: R package for causal mediation analysis.* UCLA Statistics/American Statistical Association.

Todd, J. T., & Warren Jr, W. H. (1982). Visual perception of relative mass in dynamic events. *Perception*, *11*(3), 325–335.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, *21*(9), 649–665.

Unity Technologies. (2022). *Unity.* [Online]. Available from: `https://unity.com`. (Last checked on Aug 01, 2022)

Valve Corportation. (2020). *Half-life: Alyx.* [Online]. Available from: `https://store.steampowered.com/app/546560/HalfLife_Alyx/`. (Last checked on Aug 01, 2022)

Vasta, R. (1994). Do adults perceive tilted bottles on the water-level task as rotated. In *Poster presented at the meeting of the american psychological society, washington, dc.*

Vasta, R., Knott, J. A., & Gaze, C. E. (1996). Can spatial training erase the gender differences on the water-level task? *Psychology of Women Quarterly*, *20*(4), 549–567.

Vasta, R., & Liben, L. S. (1996). The water-level task: An intriguing puzzle. *Current Directions in Psychological Science*, *5*(6), 171–177.

Vasta, R., Lightfoot, C., & Cox, B. D. (1993). Understanding gender differences on the water-level problem: The role of spatial perception. *Merrill-Palmer Quarterly (1982-)*, 391–414.

Vasta, R., Rosenberg, D., Knott, J. A., & Gaze, C. E. (1997). Experience and the water-level task revisited: Does expertise exact a price? *Psychological Science*, *8*(4), 336–339.

Vaught, G. M. (1965). The relationship of role identification and ego strength to sex differences in the rod-and-frame test. *Journal of Personality*.

Virtual Method Studio. (2022). *Obi fluids.* [Online]. Available from: `https://assetstore.unity.com/packages/tools/physics/obi-fluid-63067#reviews`. (Last checked on Aug 01, 2022)

Witkin, H. A., & Goodenough, D. R. (1981). Cognitive styles: essence and origins. field dependence and field independence. *Psychological issues*(51), 1–141.

Wu, S., Li, Y., & Kong, M. (2017). Sex and ability differences in neural strategy for piaget's water level test: An eeg study. *Perceptual and motor skills*, *124*(2), 351–365.

Zibra AI. (2022). *Zibra liquids.* [Online]. Available from: https://assetstore.unity.com/packages/tools/physics/zibra-liquids-free-201207. (Last checked on Aug 01, 2022)