How contrast affects shape perception of glossy and matte surfaces: an MTurk Study

Silan He, School of Computer Science

McGill University, Montreal

April, 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Master of Computer Science

Abstract

Though online crowdsourcing has existed since the early 2000s, human perception experiments regarding shape from shading are largely run in-person. In-person experiments permit the researchers to control for as many external factors as possible. In this thesis, we test whether Amazon Mechanical Turk or MTurk is a viable crowd-sourcing platform for shape from shading research. With the help of MTurk, we replicate a previous shape from shading experiment [16] and we re-evaluate the role of contrast in the work. We find that an increase in contrast can independently improve the perception of shape. Shape perception of glossy materials is more heavily dependent on contrast than shape perception of matte materials. In particular subjects perceive matte materials better than glossy materials in low contrast conditions. Finally, we find that MTurk can be a viable platform for human perception experiments.

Abrégé

Même si les plateformes d'approvisionnement par la foule existent depuis les années 2000, la recherche dans la perception humaine de la forme par ombre s'effectue normalement face-à-face. Le sondage en personne permet aux chercheurs de controller pour le maximum de facteurs externes. Dans ce thèse, nous testons si Amazon Mechanical Turk ou MTurk est une plateforme d'approvisionnement par la foule viable pour la recherche en perception humaine. Avec l'aide de MTurk, nous reproduisons une experimentation scientifique en perception de la forme par ombre [16]. En particulier, nous réévaluons le rôle que le contraste joue dans la perception de la forme. Nous trouvons qu'augmenter le contraste peut unilatéralement augmenter la perception de la forme. La perception des matériaux brillants est très dépendante du contraste donc les matériaux mats sont plus visible sous les conditions de faible contraste. Finalement, nous trouvons que MTurk peut être une plateforme viable pour la recherche en perception humaine.

Acknowledgements

I'd like to thank my Masters supervisor Professor Michael Langer for his endless patience, guidance and support in the development of this thesis. I am grateful to the participants of my experiments for their work. I'd like to thank my family and friends for their continued support and encouragement. This research was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant to Michael Langer.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The perception of surface shape is strongly dependent on a number of factors including and not limited to: the shape of the surface, the reflectance properties of the material, the scene illuminant direction and the viewing direction. In applications where the shape is required to be perceived accurately, it is important to choose these parameters carefully.

In this thesis, we re-examine the conclusions made in a previous shape from shading experiment by Faisman and Langer [16]. The authors studied how performance of human shape perception varied depending on a number of parameters such as light slant, surface slant and material. The authors concluded that at high light slants, participants perceive the shape of glossy materials better than matte materials because the positioning of the highlights at hilltops and valley bottoms demarcate the surface maximum. Increasing light slant led to increasing accuracy of shape perception for both matte and glossy materials.

On the other hand, we hypothesized that the effect of light slant on shape perception is being confused with the effect of increasing contrast on shape perception as light slant increases. Indeed, we believed that contrast is a confounding variable. A confounding variable is a variable one might not have accounted for and it can increase the variance or introduce bias. Due to previous work [16] not controlling for contrast, we suspect that there may have been confusion between the effect of contrast and the effect of increasing

light slant. We decided to isolate the effects of contrast and light slant on shape perception. In other words, we present an improvement to the previous experiment [16] by normalizing for contrast.

Additionally, we suggest an interactive method to calibrate the experiment dynamically for each online participant's monitor's gamma prior to a visual perception survey. Since each MTurk worker has their own computer display, we needed a way to control as much as possible for monitor gamma variances and image contrast.

The previous work's experimental subjects were college students. In this thesis, we choose to test the robustness of the previous conclusions against a novel population on Amazon Mechanical Turk or MTurk. MTurk is an online platform for crowd-sourcing studies and experiments. While the platform has been around since 2005, visual perception esperiments have historically been conducted in-person. We evaluate the reliability of the MTurk population for academic experiments through a survey of MTurk papers since 2005. We also evaluate the feasibility of running a visual perception experiment on the platform by carrying out a series of shape from shading experiments. We hypothesize that the data provided by running the experiments online on MTurk to be comparable in quality to data collected in-person.

An overview of the thesis is as follows. Chapter 2 includes background information. Chapter 3 presents the methods used for the experiments. Chapter 4 has the results of the various experiments. Chapter 5 consists of the discussion and conclusion of the thesis.

Chapter 2

Background

2.1 Mechanical Turk

Since its inception in 2005, Amazon Mechanical Turk [22] has been touted as a service for "crowd-sourcing" human intelligence tasks or HITs. The name Mechanical Turk comes from 'The Turk', a fake chess-playing automaton of the 18th century. For more than 80 years, it played and defeated many challengers such as Napoleon Bonaparte and Benjamin Franklin. It was later revealed that the automaton was actually a chess master operating the operation of the machine from a hidden compartment.

Mechanical Turk the web service allows humans to perform tasks that machines are not suited for. Mechanical Turk or MTurk became a popular platform amongst researchers of various disciplines because it helped obtain cheap data for labour intensive tasks. MTurk's Application Programming Interface or API is supported by the AWS Software Development Kit. The API allows you to programmatically distribute tasks to workers. MTurk also provides an easy to use requester user interface featuring more than 30 templates of HITs. These templates can be customized and deployed with just a few clicks. These features make MTurk accessible for all levels of programmers. Researchers or requesters post HITs for people or workers to complete. Now more than ever, with the meteoric rise of work from home in 2020, even more researchers are looking to the platform

for their surveys. Still, an unfamiliarity with the online labour markets, an uncertainty about demographic characteristics of their participants and the concerns about the data quality may delay researchers looking to adopt the platform in their research toolkit.

The platform made headlines very quickly after its inception. In 2006, a Masters thesis first put MTurk to use by requesting over 10,000 drawings of sheep [26]. Each Turker was paid 2 cents for each sheep drawn. Out of the approximate 10560 sheep submitted, 662 sheep were rejected. This makes for a rejection rate of about 6.28%. The author Koblin was intrigued and somewhat disturbed by the surprisingly overall high quality of the sheep drawings obtained by the meager 2 cent reward [26]. His work showed the potential of MTurk to provide high quality data for "very low cost" or much below minimum wage in its free market model.

In 2008, corporations started using the platform to obtain cheap feedback on their products [25]. The study established early on that the platform was a faster way to obtain data than traditional surveys. The data quality was comparable to traditional methods. At the time, there were still very few academic studies who used MTurk directly.

Since then, MTurk usage has exploded in popularity in recent years in a variety of fields. In 2005, less than 5% of psychology research studies were conducted online. In 2015, this number has risen to 50% [43]. Similarly, others predict that in coming years, nearly half of all cognitive science articles will involve online samples [6]. Searching MTurk on Google Scholar yields less than 30 results in 2005, around 500 results in 2010, 3250 in 2015 alone and more than 9000 results in 2020.

2.1.1 Demography of MTurk

In Mechanical Turk, one may directly restrict HIT participants to populations one wants to query according to diverse qualifications. For example, one may restrict their survey demographic to populations of a certain age, gender, occupation, economic status, country of residency, education level, etc. There still exists a small likelihood that the worker has lied about their qualifications on their MTurk application [31].

MTurk workers are made up of a diverse group. They vary widely in terms of age range, education levels and socio-economic strata. However, they live primarily in highly industrialized societies. Indeed, English language mastery and access to the Internet limits the potential range of MTurk workers. An ongoing demographics study [12] run since 2015 reveals that most workers live in the USA (75%), followed by India (16%), Canada (1.1%), Great Britain (0.7%), Philippines (0.35%), and Germany (0.27%).

MTurk reports having 500,000 registered workers from 190 countries. However, it is estimated that if one used the MTurk population, one sampled from a population of about 7,300 workers [44] because the more active workers crowded out the less active workers.

The US MTurk population is the largest on the platform. US Turkers are more college educated than the US general population (50% vs 25%). The US MTurk population also tended to have lower levels of annual income than the general US population. This is surprising considering the MTurk population is college educated. MTurk is not demographically representative of the broader U.S. population [42]. The US MTurk worker population is majority female at around 55% . Indeed, one should not use an MTurk sample to directly represent the broader US population as it is more representative of a typical college sample (see Fig. 2.1). If a sample of the broad US population is what one is looking for, one could reweigh their answers accordingly or try probabilistic sampling [5].

The Indian MTurk population is the second largest workforce on the platform behind the US. They are mostly male, an even larger proportion of them are college educated and they report lower levels of income.

Many use Mturk as a part time or full-time job for less than $2/hour [41]. Significantly more Indians treat MTurk as a primary source of income. Very few Indian workers participate on MTurk to ”kill time”. Over 12 % of US Turkers use the platform as their primary source of income versus over 27% of Indian Turkers. Turkers have lower income than the broad US population [23].

Since MTurk is an anonymous platform, we do not know what the exact population makeup of our subjects is. We did not query for the participant information either. Re-

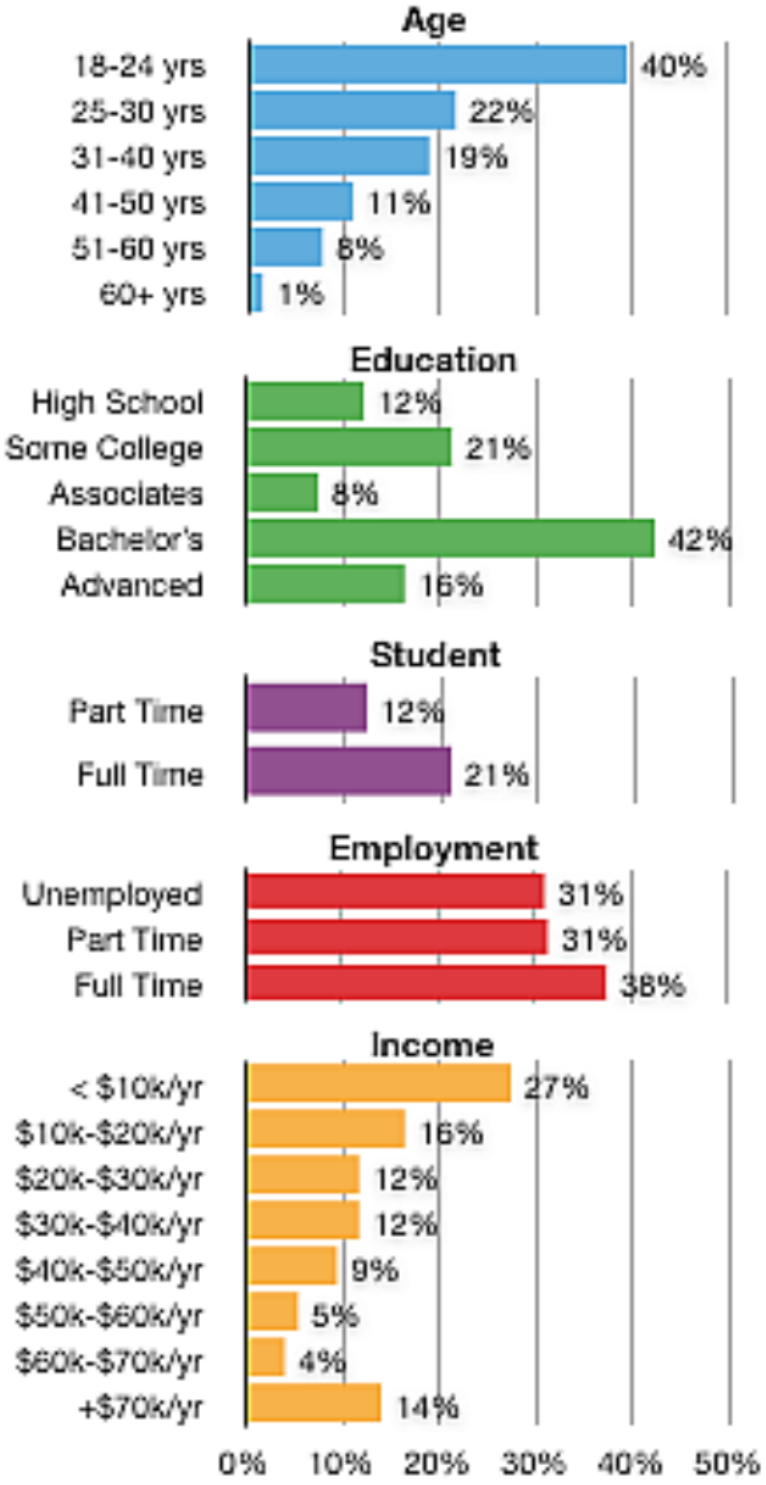


Figure 2.1: Demographics of MTurk workers [42]

searchers have been confident about the demographic stability of the MTurk population since the start of the COVID-19 pandemic [33]. Therefore, we assume that the make up of the subject population resembles the general MTurk population breakdown.

2.1.2 Data quality

The MTurk population produces reliable results for experiments and surveys at low costs [8–10, 25, 47]. MTurk may produce better data than commercial panels [49]. A commercial panel is a group of prospective research participants gathered by a company who has invested in the recruitment of people for research purposes. Still, there are concerns such as misrepresentation of the population and non-compliant responding to survey questions that may compromise the validity of research based on MTurk data [7, 21, 48]. These concerns are severe enough that some journal reviewers have essentially recommended rejecting manuscripts that used MTurk [28, 46].

Work quality on MTurk was shown to be independent of compensation rates. Compensation primarily affects the quantity but not the quality of work [32]. However, the lack of relationship between the compensation rate and quality of work was shown only with compensation below minimum wage. In fact, the data quality of India based participants is directly affected by compensation rates. The data quality of US participants, however, was not. The data of India based participants is of lesser quality than the data among US participants, even with the usage of optimal payment strategies. Optimal payment strategies refers to pay rate slightly above minimum wage for India based workers. Increasing the pay rate far above the minimum wage does not appear to further improve data quality. The motivation of MTurk workers shifted and monetary compensation is now the primary reason for working on MTurk for both US and India based participants [30].

As for the relationship between compensation and data quantity, increased pay to certain level results in greater responses for the task in question. However, pay that is too high will actually reduce demand for a task. This is due to the fact that higher pay

are linked to more complex and involved tasks [17]. The median hourly wage of MTurk workers is around $2/h while only 4% of workers earn more than $7.25 [19].

A requester's reputation matters to MTurk workers. High quality MTurk workers may avoid one's account if one has a history of rejecting lots of people, being slow to pay, or paying below platform standards. High quality workers are workers with a high approval rate (above 95%). The approval rate is the percentage of submitted tasks that the Turkers have been paid for. There are third party sites like Turkopticon that can influence a worker's interest in taking one's HIT [40]. On Turkopticon, one may rate a requester. On MTurk, one has the option of limiting workers on a HIT to only being high quality workers. For example, one can restrict HIT only to workers with above 95% approval rate [1]. MTurkers care a lot about their approval rating since it can directly influence the type of work they qualify for. These issues matter to us because we needed a criteria for accepting good data and rejecting data whose quality was not high enough.

One study reports that MTurk participants perform better on online attention checks than do subject pool participants [20]. MTurkers were shown to be more likely to pass instructional manipulation checks or IMCs than their traditional subject pool counterparts. IMCs are trick questions designed to assess participants attention to instructions. The authors argued that they would be more attentive to online instructions than traditional subject pool samples such as college students because the MTurk population encounters more online attention checks. They are more used to the patterns and have learned from them through repeated exposure. The research suggested that MTurk can reasonably be used for social science research.

Black and white decisions categorizing convenient online samples such as MTurk as good or bad ultimately harms researcher's ability to conduct research by limiting the type of samples researchers are willing to draw from [28]. Samples such as MTurk are neither good or bad, just different. Each sample has their own set of pros and cons. MTurk allows researchers to quickly obtain many responses for low cost. Conversely, the nature of anonymous online studies restricts the variables that the researcher may control. If the

pros of MTurk outweigh the cons for one's research, there is nothing wrong with using MTurk for research.

This completes our review of MTurk as a platform for carrying out research. We next turn to the main goal of the thesis project which was to carry out a perception experiment using MTurk. Specifically, we chose to replicate and expand on an earlier study of shape from shading perception [16]

2.2 Shape from shading perception

Shape from shading or SFS is a classic problem in human perception where the goal is to correctly infer the shape of a 3D scene from shading. Inferring exact shape from shading is impossible as there are infinitely many shapes, lighting conditions, and surface reflectances that can produce any given shading pattern [13]. The visual system relies on prior assumptions in order to solve the ambiguities posed by the shape from shading problem such as illumination from above [39], viewpoint from above or a globally convex surface [29]. Perhaps due to humans evolving in a solar system with a single sun, our brain also seems to prefer the 'single-light-source' assumption [39].

2.2.1 How Does Lighting Direction Affect Shape Perception of Glossy and Matte Surfaces?

Experiments in this thesis are based on the experiment from Faisman and Langer's work [16]. There, the authors present an experiment that examines local qualitative shape perception on matte and glossy surfaces. They vary the slant of the surface with the respect to the viewing direction as well as the slant of the light source. This experiment deserved a reproduction for a few reasons. One reason is the lack of gamma correction in the original study. Gamma refers to the relationship between the numerical value of a pixel and the displayed value of a pixel. Different computer monitors may have different gamma such that they display the same image in non-identical colors. Although the previous study

used a single monitor, the gamma value was not reported. When conducting perception studies, it is important to report the gamma used and to correct accordingly in order to ensure that the results are reproducible (see Sec. 2.2.2). The second reason is the previous work's authors ran their experiment with 18 subjects and we wanted to have more. The third reason is we wanted to leverage MTurk and see if a perception study offered via the platform could be a viable alternative to in-person studies.

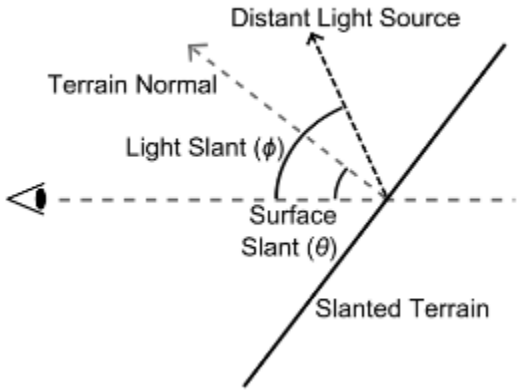


Figure 2.2: Definitions of the angles used. More details will be given in the next chapters. Indeed Figure is repeated in Chapter 3. Figure from [16].

In their work [16], the authors manipulated the slant of a surface by rotating it (see Fig. 2.2). They also manipulated the light direction and studied how these manipulations affected the patterns of shading and highlights on the surface. They vary the slant of the surface with the respect to the viewing direction as well as the slant of the light source. In turn, they analyzed how these manipulations affected shape perception. The paper reported that increasing the slant of the light source to twice that of the surface slant angle improved subjects' perception of qualitative shape but only in glossy surfaces. At high light slant angles, matte surfaces percepts were argued to be worse than those of glossy surfaces because of the positioning of highlights at the peaks and valleys of the terrain which help demaracte the surface extremas. They also found that increasing the light slant produced more consistent shape percepts than default lighting in commercial visualization software such as Matlab and Mathematica.

In this thesis, we replicated the previous study [16] but we also investigated another variable - contrast - which the authors of the previous work neglected and which we thought might be a determining factor in SFS perception.

2.2.2 Gamma Correction

Normally, when displaying stimuli for a perception experiment, the images should be presented as intended. The physical luminance should be proportional to rendered gray values. In other words, luminances should be specified exactly. This way, different experimenters would be able to replicate results. Often, images may not be displayed with the correct colors or intensity because monitors expect non-linear input for intensity. Intensity is raised to a number dubbed gamma factor. Therefore, we must gamma-correct the stimuli before displaying them. One of the motivations of the thesis is to see whether the results from Faisman and Langer's work [16] could be replicated with gamma correction.

Usually, we would determine the gamma of the monitor being used and correct for it such that the output luminance is proportional to the rendered image intensities. However, since we are presenting the perception experiment online, we cannot easily retrieve the gamma of the monitor. Preceding the experiments presented in this thesis, we present a method to figure out the gamma of a participant's monitor (see Sec. 3.1.2).

The relationship between the rendered image intensity value of a pixel and the displayed luminance can be modeled by the following power function,

$$L = D^{\gamma} \tag{2.1}$$

where L is the displayed luminance, $D \in [0, 1]$ is the rendered image intensity value and gamma is denoted as the positive constant $\gamma > 0$.

$$D = D_{linear}^{\frac{1}{\gamma}}. \tag{2.2}$$

The non-linearity can be corrected by applying the inverse relationship and is called gamma correction.

$$L = (D^{\frac{1}{\gamma}})^{\gamma} \tag{2.3}$$

With all this in mind, note that images in this document may not necessarily be shown on the reader's display with their intended intensity values.

2.2.3 Contrast

When distinguishing objects amongst other objects, contrast plays a key role. When distingushing shape of an object itself, contrast produced by shading plays a key role as well. Contrast is the difference in luminance that make different parts of an object distinguishable. Contrast is modeled by taking the difference in luminance between of the object and other objects in the same field of view. The measure of contrast used was RMS Contrast [27]:

$$C_{RMS} \equiv \frac{1}{\mu_L}(\frac{1}{N}\sum_{i=0}^{N}[L_i - \mu_L]^2)^{1/2} = \frac{\sigma_L}{\mu_L}$$

where μ_L denotes the mean luminance, σ_L denotes the standard deviation of the target's luminance, L_i denotes the target's luminance at spatial location i and N denotes the total number of spatial locations. The target refers to the object whose contrast one wishes to evaluate. In this thesis, the target corresponds to the surface in each rendered image. We are discussing contrast because we believe it to be a primary factor in shape perception. Experiment 0 and Experiment 1 revealed the effect that a slight change in contrast from the gamma correction can have on the results. We noticed that as light slant increased, contrast increased as well. We conjectured that contrast may be a confounding variable in the previous work. We do not control for contrast in the gamma corrected reproduction (Experiment 1) of Faisman and Langer's work [16]. In Experiment 2 and 3, we normalize the contrast and the luminance of the renderings in order to remove the effect of contrast from Experiment 1.

We use RMS contrast as our measure of contrast because for a random noise surface, it is a much more representative measure than something simple like Weber or Michelson contrast. Weber is used for center-surround type images. Michelson contrast is typically more suited for use in simple repetitive patterns like a sinusoid grating [37]. In random noise, it is difficult to pinpoint which luminances have the largest effect on the contrast. Indeed, absolute measures of contrast like Weber or Michelson are not appropriate for the use case of the experiments in this thesis because they are defined by the extreme values (min and max) of luminance, and one or two points of extreme darkness and brightness are not representative of the contrast of the image as a whole.

Chapter 3

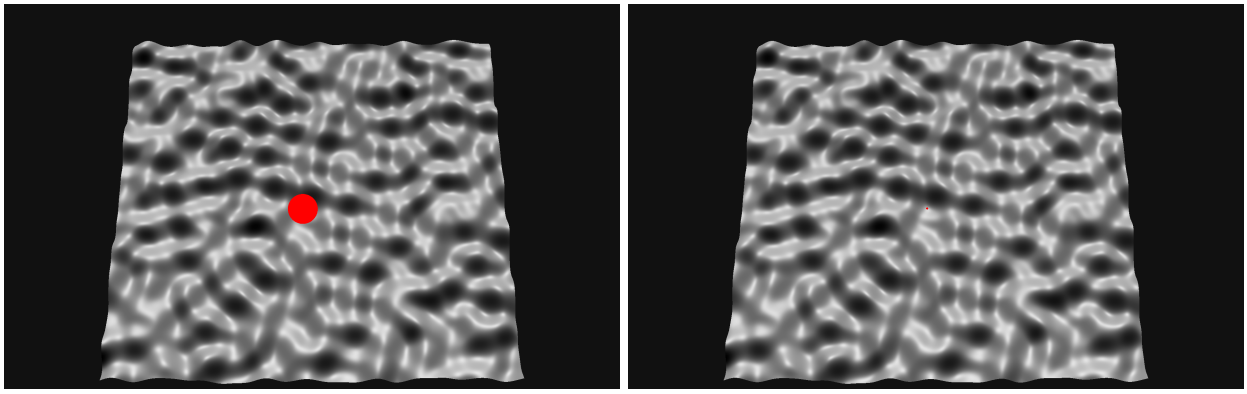
Methods

3.1 General Methods

3.1.1 Task and Stimuli

The shape from shading task requires the participant to indicate whether a probe point on the planar surface in the image is in a valley or on a hill. In each trial, we present a new surface along with a hill or valley probe point. An initial rendering with a large red sphere at the location of the probe point is used to allow the participant to make an eye movement to that general area. After 350 ms, a small red probe point with a diameter of approximately 0.2 degrees visual angle replaces the large red sphere. The participants then needs to determine whether the probe point is located on a hill or in a valley. The application presents the surface for 3.5 seconds total during which the subject must press either 'h' for hill or 'v' for valley on the keyboard. (See Fig. 3.1.) Pressing 'h' or 'v' prematurely ends the current trial and automatically starts the next one. If one fails to make one of two choices, a random choice will be made in post processing.

To each participant, we present 2 full sets of each condition for each experiment. For Experiment 0 and Experiment 1, this comes out to $88 * 2 = 176$ trials per experiment. For Experiment 2 and Experiment 3, this comes out to $80 * 2 = 160$ trials per experiment.



(a) Image shown for 350 ms. **(b)** Image shown for remaining 3150 ms.

Figure 3.1: Example of a trial. (a) is shown for 350 ms and then (b) is shown for up to 3150 ms.

Surfaces in Experiment 2 and 3 are achromatic (gray level only). The small number of trials per condition per participant has an effect on the standard error in our results which we discuss in Sec. 4.1. Each surface is randomly generated as this would reduce the likelihood of the data being affected by statistical variance caused by repeat surfaces. This meant that new surfaces had to be generated for each new participant.

Each surface is randomly generated using heights given by simplex noise which came from the simplex-rise package on npm [45]. The surface was defined using a 350×350 mesh terrain rendered in the web browser. The surfaces were generated such that each surface has about five to nine peaks per surface width. This is specified by tuning a factor which would multiply the x and y parameters of a simplex noise function. Each surface was generated in the fronto-parallel plane to the viewer. The viewing distance used was 53 cm. The camera faced the -z direction. For each surface, a probe point was placed on a convex part or in a concave part, that is, on a hill or in a valley. Again, the experiment participant determines whether that probe point is located on a hill or in a valley.

The experiment renders the surfaces with 30, 45, 60 degree rotation with respect to the normal of the fronto-parallel surface about the x axis. This is referred to as the surface slant. Recall Fig. 2.2. The surface amplitude also varies depending on the surface slant chosen. The amplitude is a chosen value that effectively represents the absolute value

maximum of the surface height value. It is a constant that multiplies the simplex noise function's result to generate random terrain. The amplitude is chosen such that the surface is fully visible. In other words, the amplitude should not be too high such that part of the surface is occluded. The occlusion contour would provide information about the local shape. The reason is that all the surfaces in this experiment have a floor-like slant (the top of the image is farther than the bottom) and, in this case, occluding contours tend to come from hills, which have an inverted U shape. So the presence of occluding contours would give information about shape. Therefore, smaller amplitudes were chosen for greater surface slants and bigger amplitudes were chosen for smaller surface slants. Each surface slant has a single different associated surface amplitude. The standard deviation of the surface heights for each surface slant were 0.1978, 0.1538 and 0.0835 for the $30°$, $45°$, $60°$ surface slant conditions respectively. The standard deviation of the surface heights is calculated by generating surfaces for each surface slant condition and their associated amplitude and taking the standard deviation of the point cloud z coordinates. The z coordinates are scaled by the 'amplitude'. It is unclear whether the change in contrast caused by different surface heights has a large effect on the perception of the different surface slant conditions. Therefore, the experiments in this thesis deal with contrast in different ways. Experiment 0 and 1 do not control for contrast. Experiment 2 and 3 deal with it by controlling for contrast by manipulating the image intensities, as will be discussed in Chapter 3.

Two reflectances are used: matte and glossy. The matte surface reflectance only has a diffuse reflectance component and no specular component. The glossy surface is composed of 0.7 diffuse and 0.3 specular (see Eq. 3.3 where $I_d = 0.7$ and $I_s = 0.3$).

$$I_{final}(x) = I_{ambient}(x) + I_{diffuse}(x) + I_{specular}(x) \tag{3.1}$$

$$I_{final} = I_{ambient}(x) + I_d\, max(n(x) \cdot \mathbf{l}, 0) + I_s(\mathbf{H} \cdot \mathbf{n})^{shininess} \tag{3.2}$$

$$I_{final} = I_{ambient}(x) + 0.7\, max(n(x) \cdot \mathbf{l}, 0) + 0.3(\mathbf{H} \cdot \mathbf{n})^{shininess} \tag{3.3}$$

$$\mathbf{H} = \frac{\mathbf{l} + \mathbf{v}}{|\mathbf{l} + \mathbf{v}|} \tag{3.4}$$

where $\mathbf{n}$ is the surface normal, $\mathbf{l}$ is the direction of the light, v is the viewer direction, $\mathbf{H}$ is the 'half vector' that defines the normal that would produce a mirror reflection, x is a point on the surface. I_{final} is the final intensity, $I_{ambient}$ is the ambient component of intensity, $I_{diffuse}$ is the diffuse component of intensity and $I_{specular}$ is the specular component of intensity. I_d and I_s are the diffuse and specular light intensities respectively.

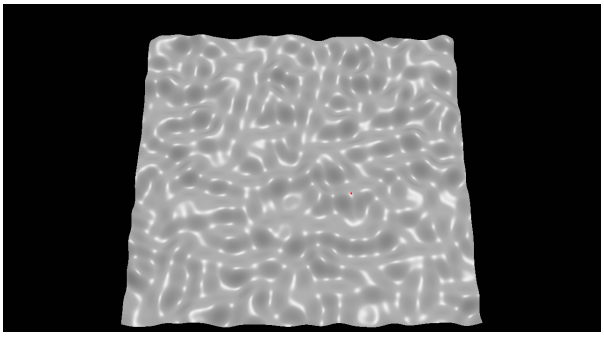
All colors mentioned correspond to RGB colors. For the matte material, 1 and 0 are used as the diffuse and specular component respectively . In addition, we used a $shininess$ exponent (also called shininess in Three.js) of 51. Three.js is the interface used to interact with WebGl in Javascript (see Sec. 3.1.3). Note that in Three.js the shininess exponent is not limited to a number between 0 and 128. Each surface was additionally rendered against a dark gray background of 0.067. All rendered values are between 0 and 1 and have no units.

Each surface will be tested under different lighting conditions, namely differently angled directional lights, the default MATLAB lighting and the Mathematica lighting.

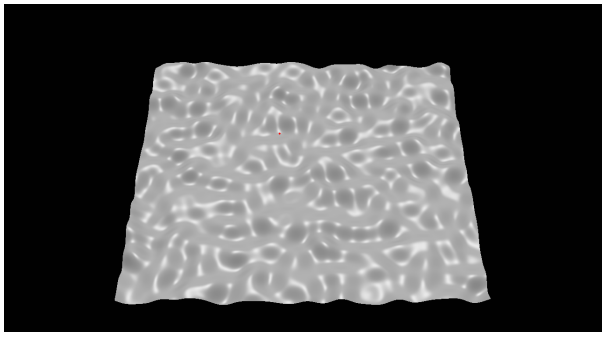
The directional lighting will consist of a directional light at infinity shining in the direction of the surface. The line of sight of the viewer to the centre of the surface is referred to as the viewing direction. For our purposes, the angle formed by the viewing direction and the light direction is dubbed light slant (see Fig. 3.3).

A different range of light slants were tested with each surface slant. We chose the same conditions as in the previous work [16], which were chosen to avoid low contrast images. Refer to Fig. 4.2 for specific tested conditions .

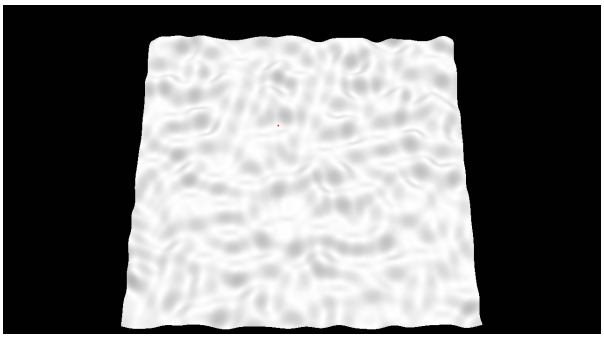
The Matlab lighting consists of the default lighting direction used by MATLAB, a directional light source from (1, 0, 1) or at 45 degrees azimuth.



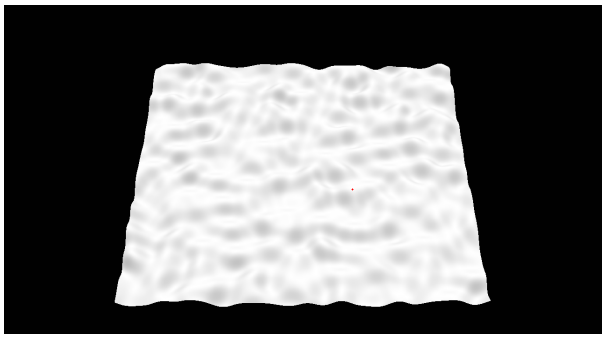
(a) 30° surface slant, 40° light slant, glossy



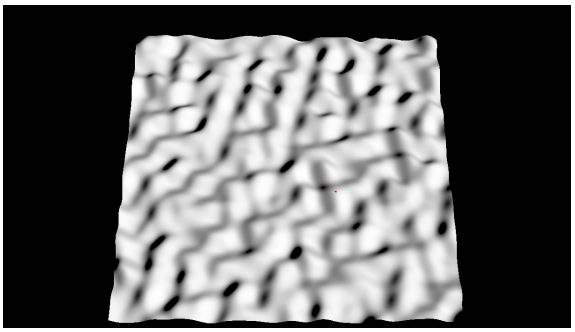
(b) 45° surface slant, 60° light slant, glossy



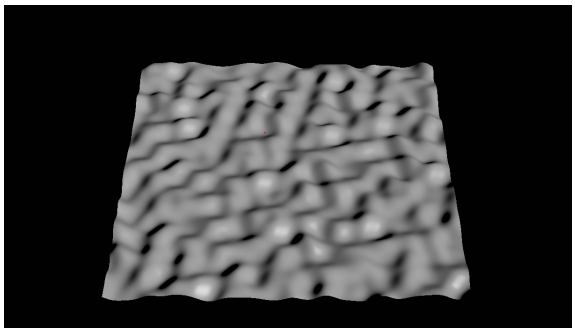
(c) 30° surface slant, 40° light slant, matte



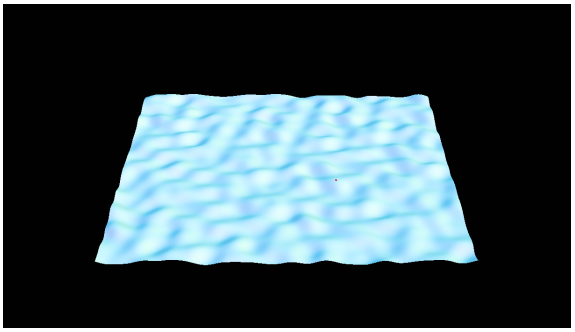
(d) 45° surface slant, 60° light slant, matte



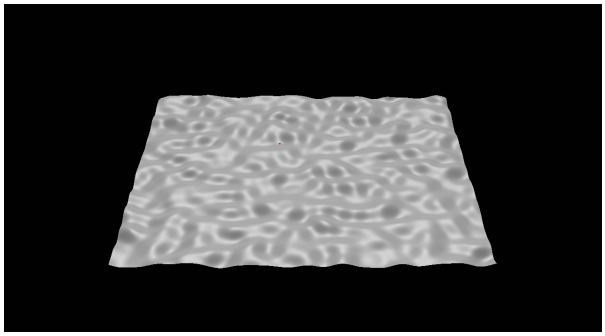
(e) 30° surface slant, Matlab, matte



(f) 45° surface slant, Matlab, glossy



(g) 60° surface slant, Mathematica, matte



(h) 60° surface slant, 100° light slant, glossy

Figure 3.2: Examples of surfaces

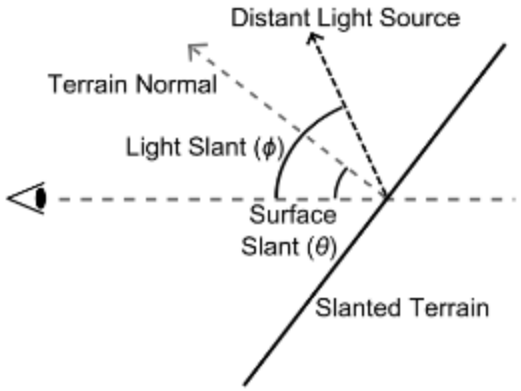


Figure 3.3: Definitions of the angles used. Figure from [16].

The Mathematica lighting consists of an ambient light component accompanied by three directional lights located at infinity. This matches the default Mathematica "automatic" lighting condition. The three diffuse light sources are coloured red, green and blue respectively. The RGB directional lights are located at infinity in the (1, 0 ,1), (1, 1, 1) and (0, 1, 1) directions respectively. (See Fig. 3.2 for examples of different conditions.)

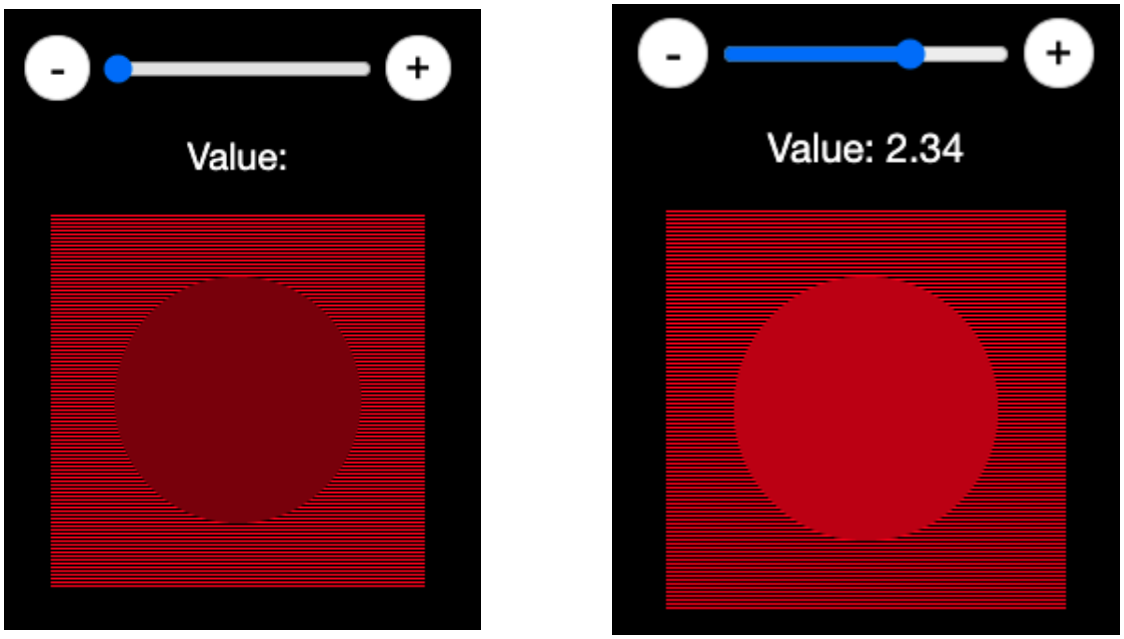
In the previous work [16], the stimuli were displayed on a 24" Apple monitor at 1920 x 120 resolution with a gamma of 2.2. No gamma correction was used. For Experiment 1 - 3, we gamma corrected according to our gamma calibration test. (See Sec. 3.1.2.)

Due to the experiment being run online, we were unable to enforce various other factors. Namely, we only recommended that the viewing distance to the monitor be roughly two monitor widths away, whereas in the lab one can control the viewing angle exactly by using a chin rest to restrain head motion. In addition, monocular viewing of the experiment with an eye patch over the non-dominant eye could not be enforced. Therefore, binocular viewing was simply allowed and monocular viewing of the experiment was not mentioned to the participants. Finally, due to pixel density variations across monitors and variations in viewing distance, the viewing angle of the stimulus was likely not exactly 17 x 13 degrees for each subject. We did not expect these factors to significantly influence results.

3.1.2 Gamma Calibration

Test used in our experiments

In order to adjust for the wide range of gammas from MTurk worker's computer monitors, certain experiment were preceded by an interactive visual gamma calibration test [3] [4]. The interactivity is also partly inspired by MacOS' own calibration test that they use for their monitors. The purpose is to estimate the gamma of the user's display. The gamma can then be corrected on the fly during the experiment.



(a) $\gamma = 1$ **(b)** $\gamma = 2.27$

Figure 3.4: Visual Interactive Gamma Calibration Test. Look at image from a distance where the the black stripes start to blend with the underlying color. Squinting can help achieve this effect as well. Participants were told to adjust the slider until the circle blends in with the striped background. Note that this may happen at different gammas for different monitors. LaTeX may not display this image as intended. Each horizontal line is defined to be 1 pixel wide in reality.

The images used for the calibration test are shown in Fig. 3.4. The images are generated as follows. The background has alternating black (0,0,0) and red (255,0,0) lines 1 pixel thick. When viewed at a distance, the background lines should blend together

and appear as a shade of red. On a linear monitor, this would correspond to looking at (128,0,0). The middle circle has an intensity that is halved compared to the background color. In Figure 3.4, a red (128,0,0) is used.

The slider at the top is used to adjust the gamma of the whole image interactively according to the formula 2.3. The idea is the participant should try to adjust the slider such that the brightness of the red circle matches that of the striped background. The $+$ and $-$ buttons help to make minute adjustments of order 0.01. The image should be viewed at a distance where the black stripes are imperceptible. Squinting may help enhance this effect. Subjects are given this hint. The calibration is repeated for each RGB channel in order to get the most accurate gamma reading possible. We take the determined gamma values and average them to a single gamma factor that we use to correct each of the generated images from the experiments.

In order to offer custom gamma correction without affecting the speed of the renderings, we extended the built-in MeshPhongMaterial's fragment shader (from THREE.Js) to incorporate gamma correction based on a gamma factor we provide as an uniform.

3.1.3 Software platform and technologies used

The JsPsych library provides a flexible Javascript framework for building a wide range of laboratory-like experiments that can be run online [11]. This library was used to register the keyboard clicks and to manage the linear progression of the experiment. It structures experiments in the form of a timeline. There is a number of trial templates that can be easily used out of the box and appended to the timeline. Trials are served in the same order in which they are added to the timeline which is a list.

The experiment website was hosted on Amazon Elastic Beanstalk or EBS which is an all-inclusive auto scaling web hosting service offered by Amazon Web Services [2] or AWS. We decided to use this service as it enables us to deploy a website in a matter of a few clicks. It is also pay as you go which is perfect for this sort of survey. Amazon DynamoDB or DDB is a NoSQL database service also offered by AWS. Since they are

both part of the same cloud ecosystem, EBS can easily send the experiment data to DDB without much additional configuration. Amazon EBS also configures auto balancing and scales one's resources according to current site traffic. Indeed, when traffic to a website hosted by EBS is too high, it will spawn additional instances of the website so as to serve all visitors with no downtime. EBS will also scale back automatically when traffic is lower, thus keeping additional costs minimal. For similar experiments, creating a new account to take advantage of free tier should cover all the costs. We only exceeded the free tier because we hosted the website for longer than we needed for demonstration purposes. Free tier on AWS includes basic usage of their services without incurring any additional cost.

NodeJs [34] and ExpressJs [36] are the platforms used to serve the experiment and host the server respectively. NodeJs is a Javascript runtime environment that allows Javascript to do things other than just making websites interactive. It allows for Javascript to be more than just a scripting language. ExpressJs is a back end web framework for NodeJs. The most important distinction here is that ExpressJs is a framework for NodeJs and not Javascript even though it has 'Js' in its name. Javascript code written for NodeJs and regular web browser Javascript runtime environments (the most basic use case of Javascript) often behaves differently and follow different rules even though the syntax is the same. The NodeJs and ExpressJs stack is an industry standard web application server that is very quick to set up. We used NodeJs version 12.8.13 because it is one of the most stable versions that Amazon Elastic Beanstalk supports out of the box. This made it easier for me to focus more on the project and less on infrastructure setup. Using a supported version of NodeJs, one can deploy a 'hello world' website to Amazon EBS for the world to see in probably less than 10 minutes.

We used Three.js [35], an application programming interface or API to WebGL [24]. WebGL is a cross-platform royalty-free web standard for a low-level 3D graphics API based on OpenGL ES, exposed to ECMAScript or Javascript via the HTML5 Canvas element. WebGL is the widely available OpenGL distribution for web applications. Three.js

and WebGL helps generate all the figures we need to serve in the experiment on the fly. We render the images on the fly as a participant loads the website in order to randomize the surfaces.

We used Node Package Manager (NPM) to manage the Javascript packages needed for the website. Namely, we used NPM to install NodeJs, ExpressJs, Three.js, AWS SDK and simplex-noise to name the core packages.

On MTurk, the 'Survey Link' HIT template was used when creating the HIT. The consent form was pasted into the survey link instructions tab that MTurkers may expand (see Sec. 3.5). The consent form was pasted once more in the experiment to ensure they will have viewed it. The link to the shape perception experiment was simply added to the template without significant change to the format. (See Fig. 3.5.)

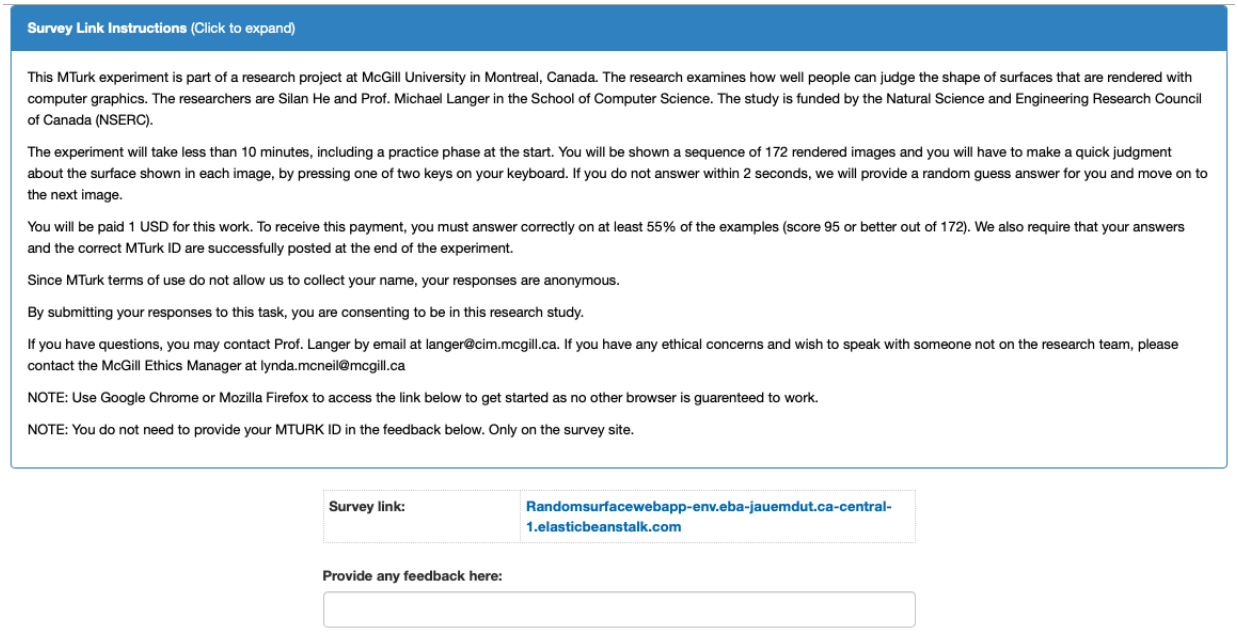


Figure 3.5: MTurk HIT Link Survey page. See Sec.3.5 for consent form.

For all experiments listed, we decided to pay participants of our survey 1 USD should they be approved. (More details can be found in subsection 3.3.)

3.1.4 Optimizations used to successfully offer the experiment online

Due to random surfaces having to be generated for each trial, it was important to properly manage the computation and stagger it effectively such that our participants are not negatively affected. In other words, we tried to remove as much lag and wait time as possible.

Each surface's point cloud is generated on the server side (meaning in NodeJs). In NodeJs, we are calling a simplex noise function to generate the point cloud. There is a GET request analogous to generating a single surface. The GET method refers to a HyperText Transfer Protocol (HTTP) method that is applied while requesting information from a particular source. In web applications, large amounts of data must be split up into smaller portions to be transferred over the web. In addition, most web browsers only wait for a server response for up to a minute. For that reason, making one large GET request, generating all 176 surfaces at once and sending the data over does not work. Instead, we make 176 different GET requests that each take a fraction of a second each.

This still results in a rather large download for the participants. Another optimization we used was to strictly send a height map of only z values. As the surface height map is generated using a predetermined size, we can infer the x y coordinates and reduce the download sizes for each request. This optimization helps reduce the size of the data transfer for each get request from 7 Mb to about 2.5 Mb uncompressed. With over 176 surfaces to generate overall, this makes a large difference in the download bandwidth required to run the experiment.

Simplex noise is used to generate the point cloud. In this study, the code for the fast simplex noise implementation came from the simplex-noise package on NPM or Node Package Manager [45]. Upon receiving the GET request, the server will generate the point cloud as well as find a hill or valley point. In order to reduce the computation time for finding the hill or valley point, a curvature estimation technique called Umbrella Curvature was used to determine suitable local maximums and local minimums [18]. Umbrella curvature uses 8 coordinates around a center point like spokes of an umbrella to produce

an estimate of the curvature at the center point. More specifically, two coordinates are taken from each of the horizontal, vertical, and the 2 diagonal axes around the center point. The curvature estimation runs in O(1) time complexity. When iterating over all suitable points of the surface, the time complexity is merely O(MN) where M x N define an area of points near the center of the surface. Suitable points were determined to be points near the center of the surface.

3.2 Experiments

3.2.1 Experiment 0

Experiment 0 is a reproduction of the experiment described in subsection 3.1. We wanted to see if the data from the previous work [16] was reproducible under slightly different conditions in an experiment offered via MTurk. In exchange for easy accessibility to a larger audience, we could not enforce monocular viewing. Viewing distance could not be enforced either. Thus, viewing distance was recommended and monocular viewing was not mentioned.

3.2.2 Experiment 1

Experiment 1 is a variation of the experiment described subsection 3.1 and 3.2.1. We wanted to see if the data from the previous paper [16] held up with a larger sample size. We also wanted to see if gamma correction for each participant's monitor would affect the results. The main differences between Experiment 0 and Experiment 1 is the presence of the gamma calibration module in Experiment 1.

3.2.3 Experiment 2

From qualitative analysis of experiment 1 surfaces conditions, we conjectured that contrast played a large role in discerning shape. In the previous work [16], the authors men-

tioned issues with low contrast which were mitigated by adjusting surface amplitudes and light source directions based on surface slant. However, they did not control for contrast.

We believe there is a theoretical confound between the effect of contrast and the $\phi = 2\theta$ argument since contrast increases when the slant of the light source ϕ increases. Experiment 2 revisits the role of contrast by controlling for it. If contrast were kept constant across all conditions, performance should theoretically be similar across all conditions. By removing the effects of contrast, we may be able to draw additional conclusions from the previous studies results. We used Eq. 3.5 to normalize all conditions to the same "target" mean and standard deviation:

$$I_{normalized} = mean(target) + \frac{std(target)}{std(I)}(I - mean(I)) \tag{3.5}$$

$$= mean(target) - \frac{std(target)}{std(I)}mean(I) + \frac{std(target)}{std(I)}I \tag{3.6}$$

where the first two terms of Eq. 3.6 can be considered the ambient light component and the third term is the same image but scaled (changing light intensity). Eq. 3.6 corresponds to Eq. 3.7 located below.

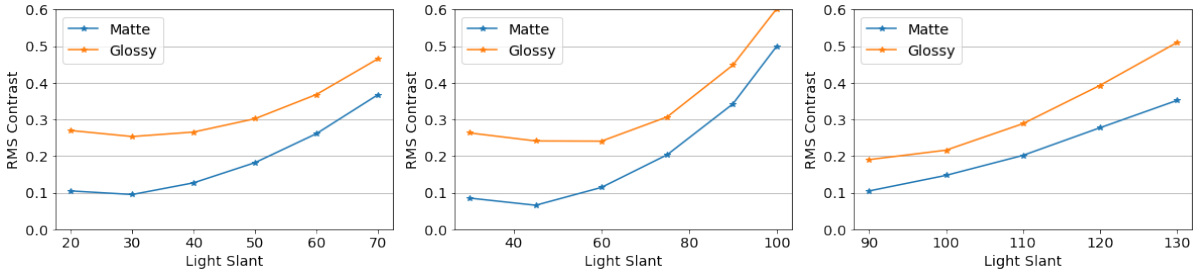
$$I_{normalized}(x, y) = I_{ambient} + cI(x, y) \tag{3.7}$$

where $I_{ambient} = mean(target) - \frac{std(target)}{std(I)}mean(I)$ and $c = \frac{std(target)}{std(I)}$. To guarantee that the ambient light term from Eq. 3.6,3.7 is positive, we require

$$\frac{std(target)}{mean(target)} < \frac{std(I)}{mean(I)} \tag{3.8}$$

That is, the RMS contrast of the target must be less than the RMS contrast of each of the rendered images. For this experiment, the target was the image with the lowest contrast. The ambient light term was required to be non-negative and we wanted all the images to have the same contrast. Plots of the RMS contrast values of each light slant condition were

used to help determine appropriate RMS contrast values to normalize to (see Fig. 3.6). We used a $mean(target)$ of 164.320 and a $std(target)$ of 10.37 which are the values from the $45°$ surface slant and $45°$ light slant condition with matte material. This condition has a target RMS contrast of 0.07. (see Fig. 3.6 where it is the lowest point of the "Matte" curve in (b).) After applying Eq. 3.5, we applied Eq. 2.2 with gamma submitted by the participant. The target RMS contrast value is the RMS contrast without any gamma correction. Gamma correction was applied post contrast normalization in each of the figures.



(a) $30°$ Surface Slant **(b)** $45°$ Surface Slant **(c)** $60°$ Surface Slant

Figure 3.6: We plotted the RMS contrast values of each light slant condition. The plots were used to help determine RMS contrast values we would normalize to for Experiment 2 and 3. Experiment 2 normalizes rendered surfaces to an RMS contrast of 0.07 which corresponds to the RMS contrast of the $45°$ surface slant and $45°$ light slant matte condition. Experiment 3 aims for RMS contrast of 0.25 which is a value that is reasonably reached by each of the curves.

The Mathematica surfaces were not considered as part of this experiment. Since the Mathematica surfaces are not grayscale, normalizing for contrast was not easily applicable.

Initially, we opted to treat the normalization in post processing but that turned out to be much too slow. Each rendering would easily take up to one second which made taking the survey very tedious. We ended up using a lookup table of pre-computed $mean(I)$ and $std(I)$ values. We computed the average luminance and average standard deviations of 100 random surfaces for each condition and used those values in the lookup table of

fragment shaders. This eliminated having to calculate the mean luminance and mean standard deviation of each image at run time.

3.2.4 Experiment 3

After running Experiment 2, we realized the contrast was too low, causing the performance on all conditions to be poor. Therefore, we had to run another experiment to fix this problem. In Experiment 3, some of the images were given a greater contrast and some were given lower contrast. In Experiment 2, all images were given lower contrast. For Experiment 3, we chose to normalize to a RMS contrast value of 0.25 (see Fig. 3.6) using Eq. 3.6. This is a much higher RMS contrast value than in Experiment 2. We felt this was a reasonable value that all surfaces from each condition could be normalized to (see Fig. 3.6). So in terms of the ambient light terms in Eq. 3.6 and Eq. 3.7, some conditions had a negative ambient term while others had a positive ambient term. We do not get negative intensities because the target RMS contrast was so low. Gamma correction was applied after normalizing for contrast in each figure's rendering process.

Like in Experiment 2, the Mathematica surfaces were not considered for this experiment.

3.3 MTurk Participants

Due to the anonymous nature of MTurk, we hypothesized that many participants would try to game the experiment which would lead to poor data. For each trial in the experiment, the participant faces a binary choice (hill or valley). In order to raise our likelihood of obtaining good data, we indicated that participants would only get paid if they score above 55 % (or slightly above chance) on our experiment. At the end of the experiment, we indicated their score. Turkers who scored below 55% should theoretically avoid submitting the task or return the task to avoid being issued a rejection. This directly affects

their approval ratings and can limit the quantity and quality of studies in which they can participate.

So when a participant scores above 55% in the experiment, they get approved and paid. In addition, we hypothesized that many participants might only pay attention for a small portion of the experiment and then just let the random choice do the rest of the experiment.

Naiveté of participants was preserved by restricting participation in the experiments to users who have never participated in an experiment run by us – including pilot studies that are not discussed in the thesis. In order to reduce nonnaïveté, we set up each batch such that MTurkers that have taken my experiments before cannot see the experiment. One can do that in MTurk by assigning a custom qualification to past workers and restricting potential workers on the new task to users that do not possess the qualification.

For each experiment, we tried to recruit 100 subjects and thus we ran experiments with batch size 100. Due to rejecting users who scored below 55% and users who tried to "game" the system, each experiment ended up with a minimum of 56 participants.

3.4 Cost of running an experiment on MTurk

Running experiments on MTurk can be very cheap. They can just as easily become very expensive. We ran several pilot studies including studies with errors, all of which cost money. MTurk takes a 20% cut of any payout. MTurk takes an additional 20% cut of any payout one makse to participants of one's surveys when one exceeds a sample request of 10 MTurkers. There is a minimum fee of $0.01 that MTurk charges for any payout. In all studies performed in this thesis, We ran experiments in batches of 100. In theory, one could run batches of 10 and save 20% on costs. We were unaware that there was different pricing for different batch sizes until we finished conducting our experiments. In the end, we spent $887.6 overall on MTurk.

If one is interested in running a highly custom experiment, it is likely one will also have to spend on server costs for their custom experiment. We recommend using free cloud credits that various providers such as Google or AWS provide to new account holders in order to cut costs.

3.5 Appendix: Waiver presented prior to experiment

Here is the text of the waiver that each MTurk worker needed to read and acknowledge prior to the experiment:

This MTurk experiment is part of a research project at McGill University in Montreal, Canada. The research examines how well people can judge the shape of surfaces that are rendered with computer graphics. The researchers are Silan He and Prof. Michael Langer in the School of Computer Science. The study is funded by the Natural Science and Engineering Research Council of Canada (NSERC).

The experiment will take less than 10 minutes, including a practice phase at the start. You will be shown a sequence of 160 rendered images and you will have to make a quick judgment about the surface shown in each image, by pressing one of two keys on your keyboard. If you do not answer within 2 seconds, we will provide a random guess answer for you and move on to the next image.

You will be paid 1 USD for this work. To receive this payment, you must answer correctly on at least 55% of the examples (score 88 or better out of 160). We also require that your answers and the correct MTurk ID are successfully posted at the end of the experiment.

Since MTurk terms of use do not allow us to collect your name, your responses are anonymous.

By submitting your responses to this task, you are consenting to be in this research study.

If you have questions, you may contact Prof. Langer by email at langer@cim.mcgill.ca. If you have any ethical concerns and wish to speak with someone not on the research team, please contact the McGill Ethics Manager at lynda.mcneil@mcgill.ca.

Chapter 4

Results

In Sec. 4.1, we preface the results with comments on the standard error which is represented in Fig. 4.2 by the horizontal bars directly above and below a data point. This is followed by the shape from shading experiment results in Sec. 4.2 - 4.6. Finally, in Sec. 4.7 we will detail observations and results on MTurk as a viable platform for perception research.

4.1 Comments on Standard Error

In the previous work [16], Faisman did not report how many trials he ran *per subject per condition*. For Experiment 0 and 1, we ran 176 trials to account for the hypothesized low attention spans of MTurk users. For the same reason, we chose a set of 160 trials for Experiment 2 and 3. We tried to avoid making the task tedious in order to get higher quality data. This means that we only show 4 images *per condition per subject*. As 4 images total were shown for each condition, each participant could only score 0%, 25%, 50%, 75% or 100%. Having only 4 trials per condition per participant normally leads to large standard deviation. In this case, the sample means x_i for subject i in a given condition would normally be poor approximations to the true population (all subjects) mean for

that condition. For that reason that we might expect the estimate of the standard deviation *over subjects and for each condition* to be large.

We counter this effect by surveying a large number of subjects n in our sample. (Notice the denominator of Equation 4.2.) In all our experiments, $n > 56$. Basically, the number of trials per condition per experimental subject is small but the number of subjects is large in our experiments. This ensures the standard error of the mean SE is small. Specifically, per condition, we use an unbiased estimate of the standard deviation of the x_i's:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \tag{4.1}$$

where x_i is percent correctness of participant i, n is the number of subjects and $\bar{x}$ is the mean over all i subjects for that condition. The standard error of the mean is defined:

$$SE = \frac{s_x}{\sqrt{n}} \tag{4.2}$$

and is shown in the result graphs in Fig. 4.2, namely the error bars per condition show $\pm$ SE. The error bars are quite tight. For this reason, we will not delve into statistics when analyzing the data. Instead, most of our observations will be high level and qualitative.

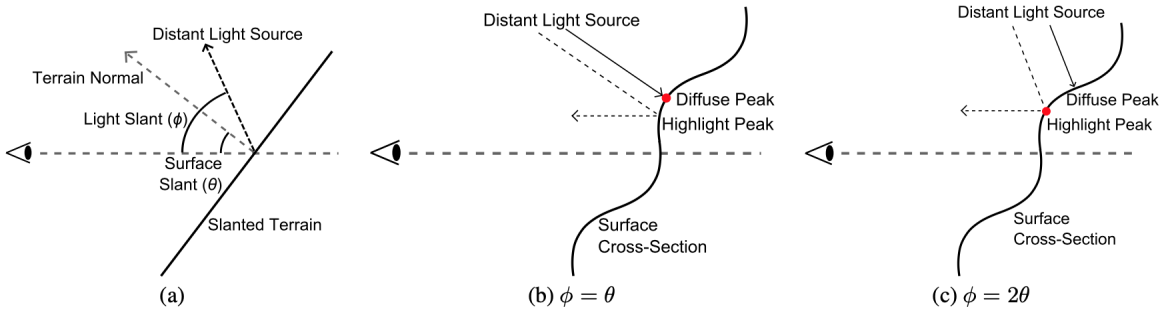


Figure 4.1: (a) Definitions of the angles used. (b,c) Typical locations of the peak components of the diffuse and highlight lighting components for two configurations of the light slant (ϕ) given a constant surface slant (θ). The case (b) demonstrates $\phi = \theta$ whereas (c) demonstrates $\phi = 2\theta$. Red dot indicates the top of a hill, which is a candidate probe location in the experiment. Figure from [16].

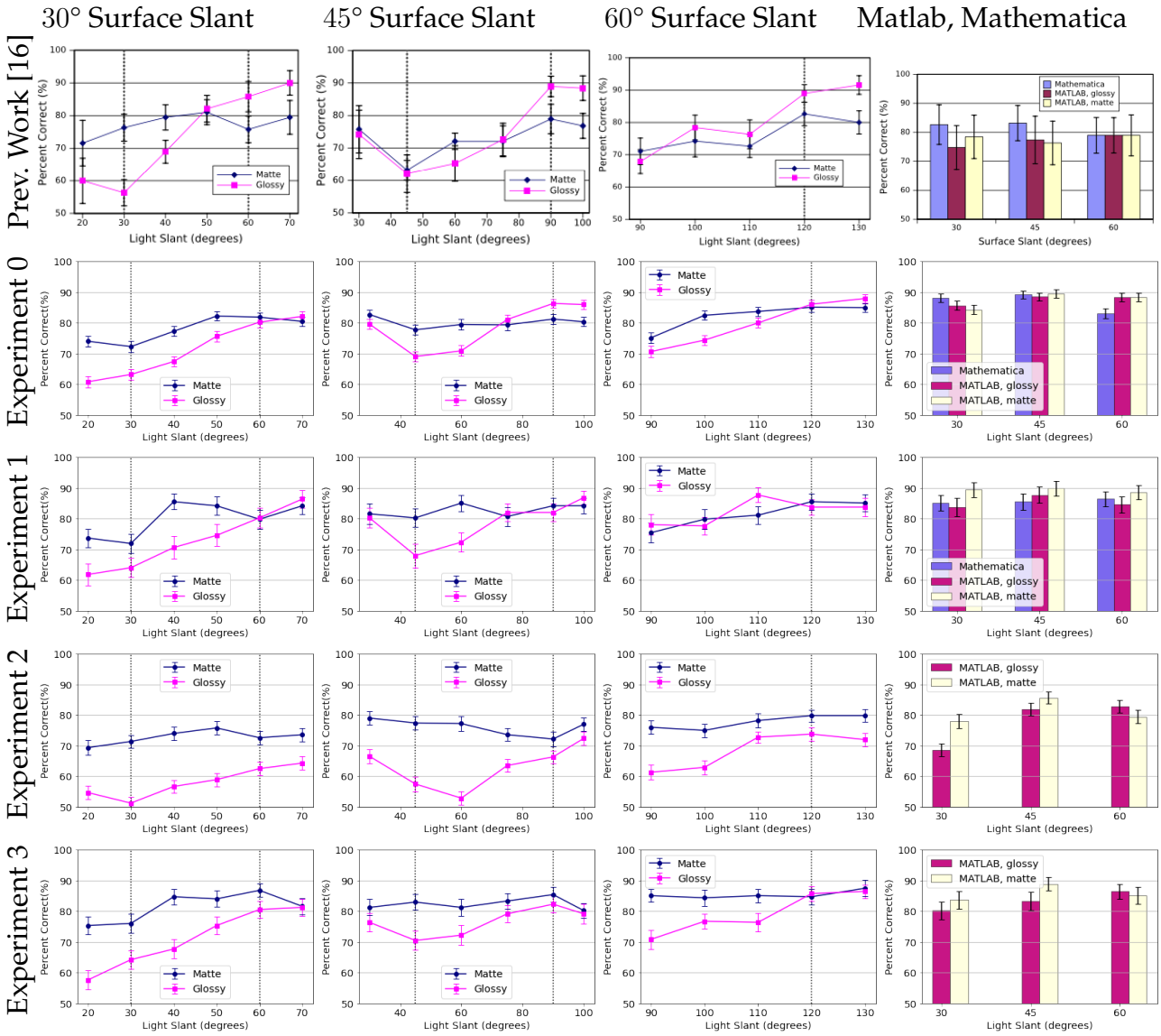


Figure 4.2: Each row shows the results one of the experiments. The first row shows the results from [16] for the reader's convenience. Within each data plot, the first vertical dotted line marks the $\phi = \theta$ condition and the second vertical dotted line marks the $\phi = 2\theta$ condition. Error bars show standard error (SE).

4.2 Results from [16]

Before presenting the results of our experiments, we first review those of [16]. Fig. 4.1 is taken from that paper. To understand these results, we review some of the basic arguments and observations made in that paper.

For rendered images used in the experiments in the previous work and this thesis, the location of the high intensity peaks from the diffuse component and from the glossy component occur at different locations. The locations where the peaks occur depend on the light slant. Fig. 4.1 (b)(c) shows the location of the diffuse peaks and the glossy peaks when $\phi = \theta$ and when $\phi = 2\theta$. We define ϕ to be the light slant and θ to be the surface slant. As shown in Fig. 4.1(b) at $\phi = \theta$, the *diffuse* peak is located at the highest point of the hills and the lowest point of valleys. As shown in Fig. 4.1(c), at $\phi = 2\theta$, the *highlight* peak is located at the highest point of the hill. The intensities also peak at the bottom of the valleys because the surface normal is parallel to the intensity peak on the hills.

The data from the previously published work is reproduced in Fig. 4.2 first row. As previously reported, once $\phi = 2\theta$ is reached, the glossy condition seems to produce consistently higher performance [16]. As described above, $\phi = 2\theta$ corresponds to conditions where the light slant yields highlights appear exactly at the tops of hills and the bottoms of valleys. At these high light slants, the peaks on diffuse surfaces are located on the oblique part of the surface above hilltops and are foreshortened, which provides an effective shape cue [15]. In addition, the previous work showed that increasing the light slant lead to better shape perception than in the Matlab and Mathematica built-in lighting conditions.

Finally, the previous work mentions one last factor that may play a role in perception of shape, namely at high light slant ($\phi = 2\theta$) diffuse shading becomes linear [38]. With linear shading, intensity maxima appear where the surface slope is the greatest with respect to the terrain, instead of at the hilltops and valley bottoms. Linear shading generally produces lower spatial frequencies in image intensity We will discuss linear shading again shortly, in the context of the normalization of the image intensities in Experiments 2 and 3. For now, we begin our discussion of the results with Experiment 0.

4.3 Experiment 0

In the previous work (see Fig. 4.2 first row), in-lab subjects perform substantially better on the glossy conditions than on the matte conditions at high light slants ($\phi = 2\theta$). In Experiment 0 (see Fig. 4.2 second row) , MTurk subjects perform marginally better on the glossy conditions than on the matte conditions at high light slants ($\phi = 2\theta$).

The MTurk subjects perform much better at the Mathematica and Matlab conditions than the in-lab subjects of the previous work. Unlike the previous results, increasing the light slant does not always lead to better shape perception when compared to the Mathematica or Matlab lighting conditions.

4.4 Experiment 1

In Experiment 1, we are reproducing the previous work's experiment but with gamma correction. At high light slants, subjects perform similar on the glossy conditions compared to the matte conditions. The data from Experiment 0 and Expeirment 1 look nearly identical. Unlike the previous results, increasing the light slant does not improve shape perception substantially compared to the Mathematica or Matlab lighting conditions.

4.5 Experiment 2

Recall that the main motivation for Experiments 2 and 3 is to examine a potential theoretical confound between the effect of contrast and the $\phi = 2\theta$ effect, namely contrast also increases when ϕ increases. Experiment 2 removes the contrast factor that may have led to the Experiment 1 results.

In Experiment 2, the Turkers perform worse in all conditions but are still able to discern shape. The performance on Experiment 2 is low overall because the surfaces have low contrast (see Fig. 4.3). Performance for the glossy surfaces was much worse than before when compared to matte surfaces. The absolute values of the performance for the

glossy conditions is rather close to chance in certain conditions, namely when either the surface slant is low ($30°$) or the light slant is low.

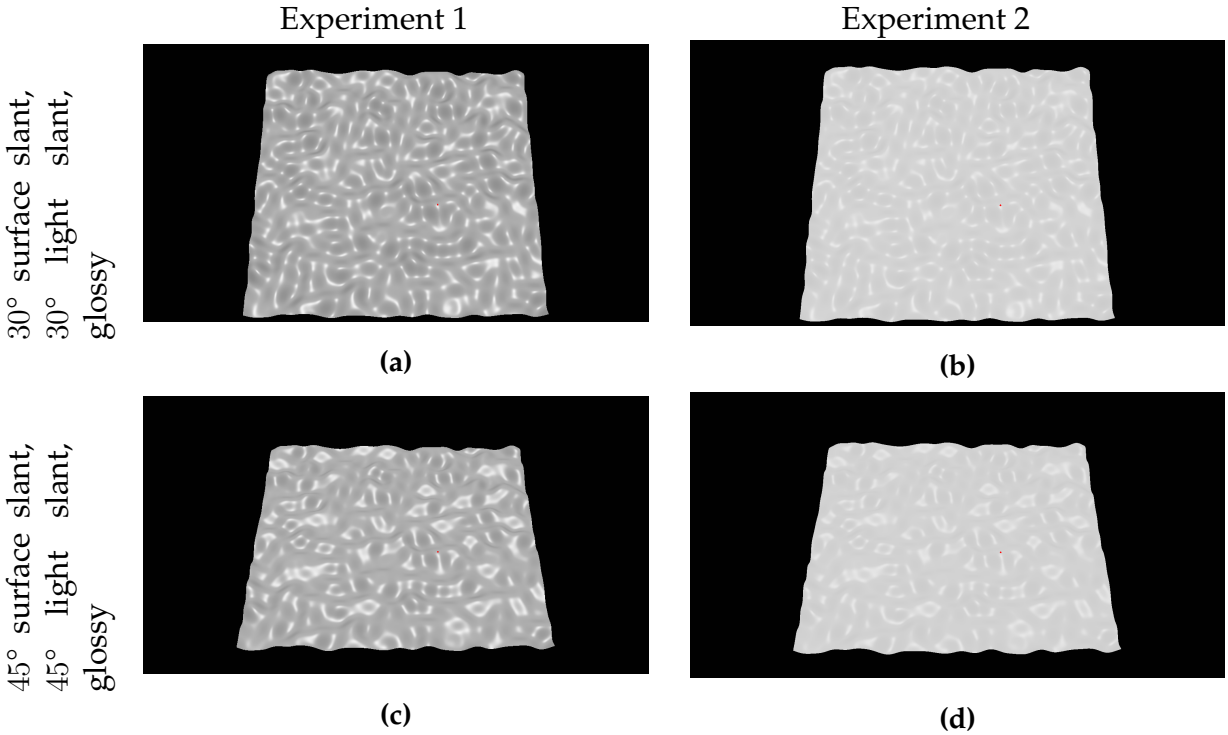


Figure 4.3: Experiment 2 surfaces have very low contrast compared to Experiment 1 surfaces

For the $60°$ surface slant matte condition, all different light slant conditions render the same image in Experiment 2 where images are normalized to the same contrast (see Fig. 4.4). This is because at high light slant, the diffuse shading becomes linear [38]. See the paper for the derivation. Specifically, the usual rendering model for matte surface illuminated from direction L is

$$I(x, y) = \mathbf{N}(x, y) \cdot \mathbf{L}$$

where $\mathbf{N}(\mathbf{x}, \mathbf{y})$ is the local surface normal, which depends non-linearly on the height $Z(x, y)$ of the terrain. Under certain conditions, this model can be approximated by a

linear model

$$I(x, y) = I_0 + I_1 \frac{\partial Z}{\partial x} + I_2 \frac{\partial Z}{\partial y}$$

where $\frac{\partial Z}{\partial x}$ and $\frac{\partial Z}{\partial y}$ are the slopes of the surface in the x and y directions. This linear shading model is accurate when the magnitudes of these slopes are small (low relief) and when the surface is illuminated from an oblique angle.

In particular, observe that normalizing the images by setting the mean and contrast to some target value will just lead to another linear model, which has some target mean instead of I_0 and the constants I_1 and I_2 will be rescaled such that the target contrast is obtained.

When comparing identical surfaces from Experiment 1 to their contrast normalized counterpart in Experiment 2, we see that the spatial pattern of the shading is actually identical across all the light slant conditions. The contrast of the shading in Experiment 1 increases as light slant increases but the spatial pattern of the shading stays the same (see Fig 4.4). Thus we may conclude, in Experiment 1, the improvement of shape perception for the $60°$ surface slant matte conditions as light slant increases seems to have come purely from the increase in contrast.

The shape from shading performance of subjects on glossy sees a large drop off when comparing Experiment 1 and Experiment 2. Shape from shading of subjects on matte sees a smaller performance drop off when comparing Experiment 1 and 2. Indeed, perception of shape from shading in glossy conditions is more heavily influenced by low contrast than the perception of shape from shading in matte conditions. The lower performance in the glossy conditions could be directly attributed to the normalization of contrast in Experiment 2, namely to a contrast value lower than in Experiment 1 (see Fig. 3.6 and note that contrast is lower in the right column (Exp. 2)). In Experiment 3, we correct the low contrast performance issue by normalizing the images to a higher RMS contrast value.

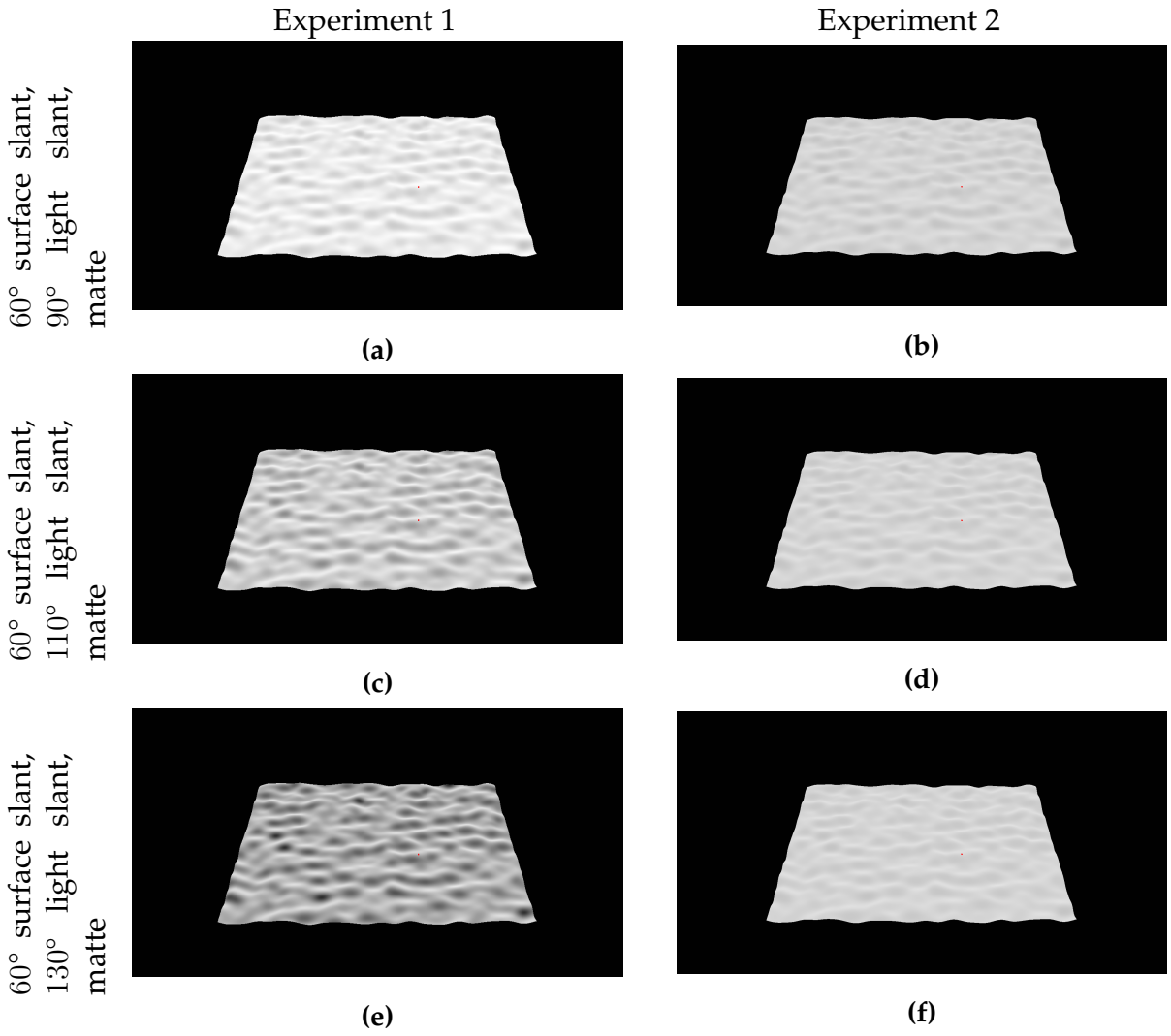


Figure 4.4: Experiment 2 surfaces are all identical. At high light slant conditions, the increase in the people's performance from Experiment 1 may come directly from an increase in contrast for the $60°$ surface slant conditions.

4.6 Experiment 3

Going from Experiment 2 to Experiment 3 (see last 2 rows of Fig. 4.2), the percent correctness of shape perception for all conditions improved. Indeed, increasing contrast can directly lead to improvement in shape perception. This is in line with the expectation that contrast plays an especially large role in shape perception.

The performance in perceiving the $60°$ surface slant matte conditions is flat with regards to light slant since this is a linear shading situation and so the normalized images for this condition are all roughly the same. (see Fig. 4.5 right column). Similarly to our linear shading argument in Experiment 2, we conclude that for Experiment 1, the increase in performance in matte conditions for $60°$ surface slant conditions and across the (high) light slant conditions tested was purely due to contrast. The increase in performance due to contrast can also be seen for the $30°$ surface slant conditions and the $45°$ surface slant conditions (see Fig. 4.2 data to the right and including second vertical dotted line).

In summary, contrast has a negative effect on shape perception when there is not enough of it. An increase in contrast can independently lead to an enhancement in the perception of matte and glossy surfaces. An increase in light slant does lead to an improvement in shape perception even when normalizing contrast.

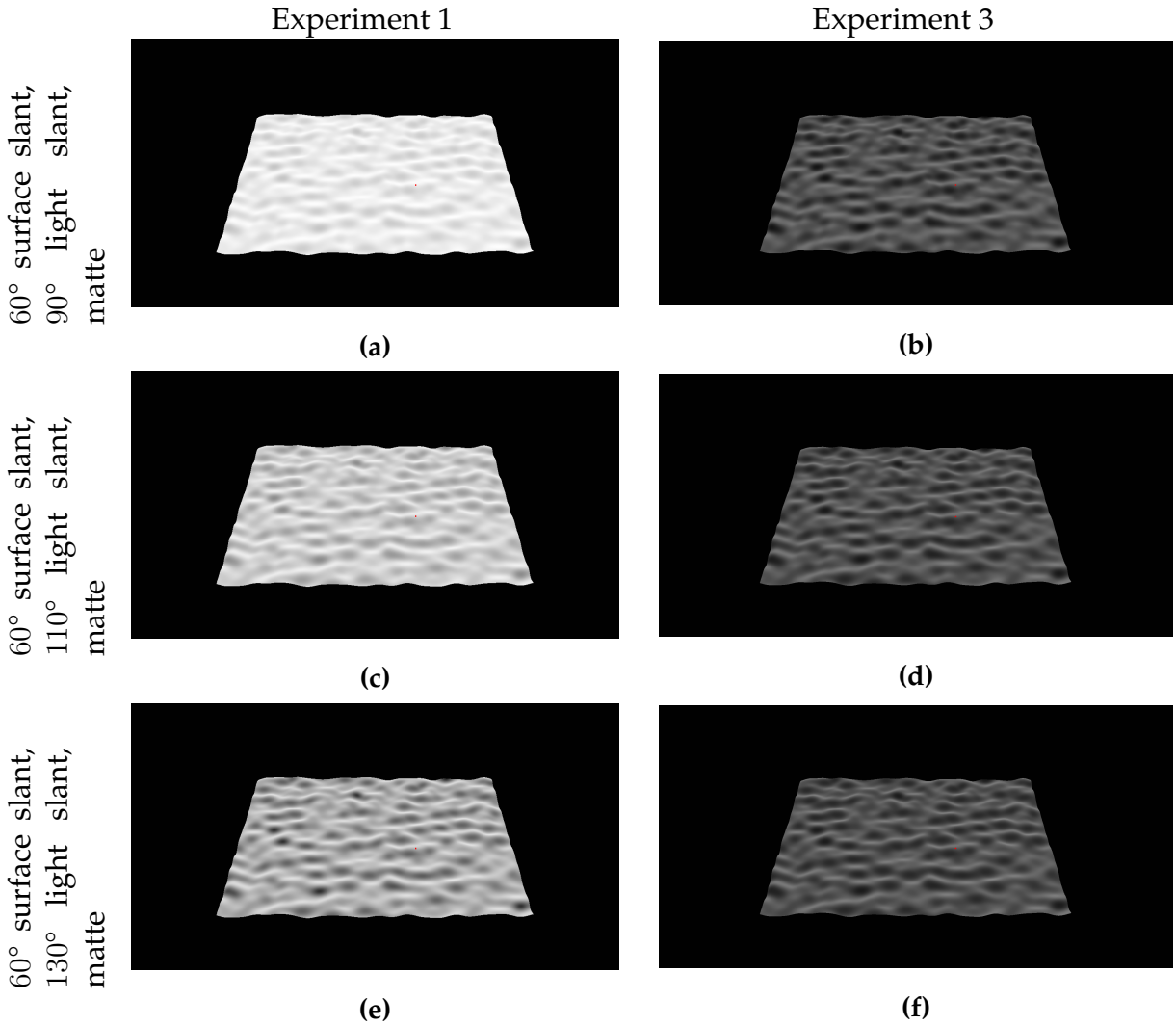


Figure 4.5: Experiment 1 and Experiment 3 figures.

4.7 MTurk issues

In addition to exploring contrast effects on shading, another motivation of this work was to examine how well MTurk could be used. Here we discuss some of the issues.

4.7.1 Data Quality

We decided to only consider all data for people who scored more of equal to 55% overall ('Approvals' in Table 4.1). If we consider all the MTurk data (including 'Rejections' and 'Incomplete' in Table 4.1), it only shifts the data points down. When considering the whole data set, the data quality from [16] was higher when compared to the data to the MTurk version of the experiment. When only considering the data for participants who scored more than 55% overall on this 50-50 task, the MTurk subjects outperform the in-person subjects.

Batch Details	Approvals	Rejections	Incomplete
Exp 0, 1	52	35	13
Exp 0, 2	41	44	15
Exp 0, 3	48	49	24
Exp 0, 4	47	37	24
Exp 1, 1	56	29	15
Exp 2, 1	67	26	7
Exp 2, 2	66	25	9
Exp 3, 1	66	27	7
Total	443	272	114
Percent (%)	53	33	14

Table 4.1: Batch data. Starting Experiment 1, batch sizes of 100 were used. Two batches were run for Experiment 2 to confirm a statistical variation. After including the second batch of data, the statistical variation disappeared. 'Approvals' scored more of equal to 55% and 'Rejections' scored below 55%. 'Incomplete' refers to participants who submitted the task on MTurk but did not fully complete the experiment on my website. These people tried to 'game' the system. Their data were not included.

Throughout this experiment, more than 30 % of eligible participant data were found to be below the arbitrary 55% threshold (refer 'Rejections' column in Table 4.1). The overall acceptance rate of these participants hovers just below 50% for all tasks given (see Fig. 4.6(b)). This suggests that these people simply could not do the task or that they were guessing.

Due to the experiment not being hosted on MTurk itself but rather being hosted on my own website, certain users could be expected to try to game the experiment and directly submit the survey on MTurk but not have completed the experiment which is hosted on the separate website. This corresponds to the 15% of data which was dropped because the subjects were not found within the database. These 15% consists of participants did not complete the survey but submitted the MTurk task regardless, likely hoping to game the system. They are reflected in Table. 4.1 as the 'Incomplete' column. Starting Experiment 1, batch size of 100 were used. The core experiment remained the same throughout. Due to the large amount of rejections, each experiment ended up with a minimum of 56 subjects.

We paid each participant 1 USD for a task that takes less than 10 minutes. This time includes time required to read the instructions and time needed to download the data. This should bring the compensation of the experiment close to the minimum federal US wage. By this train of thought, one should expect that the data is of high quality. Considering this, we can choose to ignore the arbitrary 55% threshold and consider the data as a whole. Due to the small sample size of 18 subjects in the previous work, our larger sample size could be a more accurate representation of perception abilities of the population as a whole. It remains to be seen why about 30% of participants 'failed' to score above 55%. Most researchers approve more than 90% of survey takers on MTurk. In fact, less than 10% of requesters reject more than 10% of survey takers on MTurk [14].

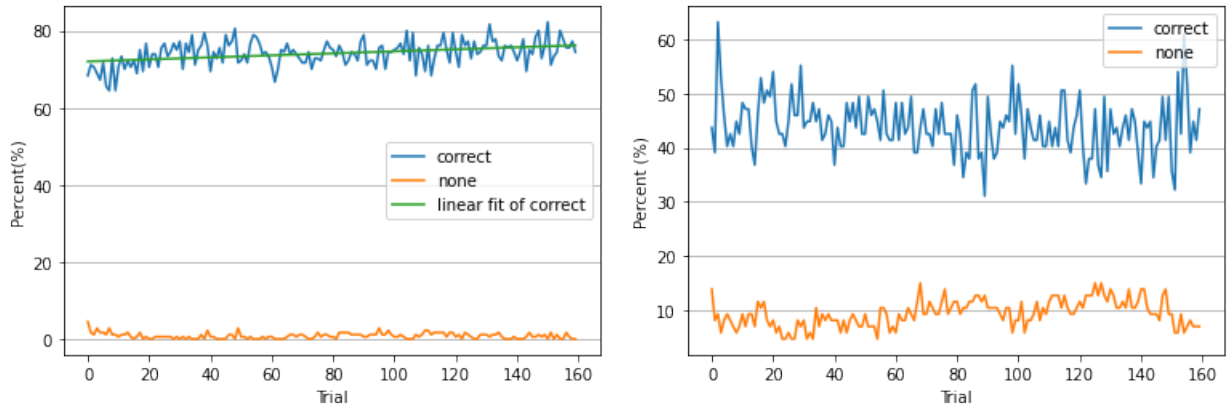
We presented the MTurker's percent correctness to the survey website as soon as they completed the survey. This would allow them to return the task on MTurk if they wished preserve their approval ratings. Still, quite a large portion of the submitted data did not satisfy the 55% correctness threshold that warrants payment. These people submitted

their task on MTurk even though it was clearly indicated thrice that only people who scored at least slightly above chance would be paid. We received a total of 41 emails over the course of all the experiments from people complaining about the passing threshold even though they agreed to the waiver. Most of the 41 came from our pilot experiments. In order to minimize the number of such emails in the actual experiments i.e. reported in this thesis, we indicated the 55% passing criteria in the MTurk 'description' of the experiment on the MTurk website. This would ensure that there is visual evidence that this criterion was mentioned even after we take down the experiment website. We believe this led to fewer complaints in the latter studies. All of the 41 people responded directly via MTurk and not using the emails provided in the waiver (from Sec. 3.5).

4.7.2 Feedback from MTurkers

For all my studies, we asked for feedback on the experiment. Some complained that a survey completion code should be provided since it is standard with surveys conducted outside of MTurk. One individual even thought our experiment might be a scam because of the lack of survey completion code. For future reference, even though a survey code was not required to properly acknowledge and pay participants, it might be better to include one.

Through the feedback from the pilot studies, we were able to clean up a lot of the issues with the experiment and how it was conducted. We would recommended researchers considering MTurk to run small batches on MTurk purely for feedback on the study as well. As the feedback loop of running an experiment on MTurk is quite short, this allows researchers to quickly improve their studies. MTurkers in the pilot studies were happy to provide constructive feedback on the survey.



(a) Participants above 55% correctness overall **(b)** Participants below 55% correctness overall

Figure 4.6: (a) Correctness average increases during experiment for Experiment 1 - 3. When fitting the correct data points in (a) with a linear fit, we get a line with $m = 0.026$ and $b = 71.97$. The 'none' line represents how many percent of people failed to click either h or v in response to the stimulus. The 'correct' line represents the percentage of people who got the right answer.

4.7.3 MTurkers learn

Over the course of the experiment, participants of the experiment actually improved their percent correctness (see Fig. 4.6). This shows that on average most participants paid enough attention and even learned to perform better without any feedback. It may also be interesting to observe how much participants improve over the course of a larger sample.

Chapter 5

Discussion and Conclusion

We found that trends in the directional light conditions from the previous work [16] held in Experiment 0, which was our MTurk replication of the previous work. Unlike the subjects in the previous work, the Turkers performed much better in the Matlab and Mathematica conditions thus putting some doubt on whether increasing light slant is better than the Matlab and Mathematica built in lighting conditions as was previously concluded. Experiment 1, the variant of Experiment 0 with gamma calibration, generally reproduced the trends in the directional lighting conditions from Experiment 0 and the previous work as well. Once more, the Matlab and Mathematica perform just as well as the directional lighting conditions. The perception of glossy conditions was not better than the perception of matte conditions at high light slants ($\phi = 2\theta$) where shading is linear.

As performance between Experiment 1 and Experiment 3 is quite comparable and performance on Experiment 2 is lower than the former, we conclude that contrast is important to shape perception if there is not have enough of it. For the $60°$ surface slant, directional lighting and matte material condition, we showed that the improvement in shape from shading can be directly attributed to an increase in contrast. When we controlled for contrast in Experiment 2 and 3, every light slant condition for the $60°$ surface slant matte material condition rendered the same surface because diffuse shading becomes linear at

high light slant [38]. Since each of the surfaces rendered were the same, the perception performance was flat across these light slant conditions. In Experiment 1, we see an increase in shape from shading performance as light slant and contrast increased for the $60°$ surface slant, matte material condition. It is therefore possible to improve shape perception purely by increasing image contrast. Contrast seems to play an especially large role in glossy material conditions as the reduction in contrast caused a massive overall dip in shape from shading performance in Experiment 2. Matte materials are much more reliable in low contrast environments.

Past some threshold of contrast, the other factors stated in the previous work such as the location of intensity maxima and the location of the highlights seem to play a larger role in shape perception. It would be interesting to verify this claim by running the same experiment at a larger variety of contrast levels and seeing how the performance changes for the tasks. It would also be intriguing to find out where performance begins to suffer at higher light slant conditions.

In terms of MTurk, we conclude that running a human visual perception experiment online can be viable if one takes the proper precautions. Amongst these precautions are proper pay, a robust testing environment that enables one to easily identify people who try to game the system and per subject gamma correction. This thesis benefited from the binary nature of the trials to easily identify poor data. We recommend running studies where the questions posed to the subjects have right and wrong answers.

Bibliography

[1] Amazon mechanical turk. https://www.mturk.com. Accessed: 2020-09-03.

[2] Amazon web services. https://aws.amazon.com. Accessed: 2020-09-03.

[3] Gamma calibration. http://www.lagom.nl/lcd-test/gamma_calibration.php. Accessed: 2020-01-25.

[4] Monitor calibration for photography. https://www.cambridgeincolour.com/tutorials/monitor-calibration.htm. Accessed: 2020-01-25.

[5] Probabilistic sample - an overview. https://www.sciencedirect.com/topics/computer-science/probability-sample. Accessed: 2021-03-8.

[6] ANDERSON, C., ALLEN, J., PLANTE, C., QUIGLEY-MCBRIDE, A., LOVETT, A., AND ROKKUM, J. The mturkification of social and personality psychology. Personality and Social Psychology Bulletin 45 (10 2018), 014616721879882.

[7] BARENDS, A., AND DE VRIES, R. Noncompliant responding: Comparing exclusion criteria in mturk personality research to improve data quality. Personality and Individual Differences 143 (06 2019), 84–89.

[8] BEHREND, T., SHAREK, D., MEADE, A., AND WIEBE, E. The viability of crowdsourcing for survey research. behavior research methods, 43, 800-813. Behavior research methods 43 (03 2011), 800–13.

[9] BUHRMESTER, M. D., KWANG, T., AND GOSLING, S. Amazon's mechanical turk. Perspectives on Psychological Science 6 (2011), 3 – 5.

[10] CHANDLER, J., MUELLER, P., AND PAOLACCI, G. Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. Behavior research methods 46 (07 2013).

[11] DE LEEUW, J. jspsych. URL:http://docs.jspsych.org/, 2016. Accessed: 2020-09-03.

[12] DIFALLAH, D., FILATOVA, E., AND IPEIROTIS, P. Demographics and dynamics of mechanical turk workers. WSDM '18: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (02 2018), 135–143.

[13] D'ZMURA, M. Shading Ambiguity: Reflectance and Illumination. MIT Press, 1991, ch. 2, pp. 187–207. Editors: M. Landy and J. A. Movshon. Chapter 2: Shading Ambiguity: Reflectance and Illumination.

[14] EDELMAN, J. Should i reject this mturk worker? how to fairly identify low-quality research participants. CloudResearch website, 08 2020. URL: https://www.cloudresearch.com/resources/blog/fairly-reject-low-quality-mturk-worker/.

[15] FAISMAN, A., AND LANGER, M. Qualitative shape from shading, highlights, and mirror reflections. Journal of vision 13 (04 2013).

[16] FAISMAN, A., AND LANGER, M. S. How does lighting direction affect shape perception of glossy and matte surfaces? In Proceedings of the ACM Symposium on Applied Perception (New York, NY, USA, 2013), SAP '13, Association for Computing Machinery, p. 9–14.

[17] FARADANI, S., HARTMANN, B., AND IPEIROTIS, P. What's the right price? pricing tasks for finishing on time. In Papers from the 2011 AAAI Workshop (01 2010), The organization, Human Computation.

[18] FOORGINEJAD, A., AND KHALILI, K. Umbrella curvature: A new curvature estimation method for point clouds. Procedia Technology 12 (12 2014), 347–352.

[19] HARA, K., ADAMS, A., MILLAND, K., SAVAGE, S., CALLISON-BURCH, C., AND BIGHAM, J. P. A data-driven analysis of workers' earnings on amazon mechanical turk. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (2018).

[20] HAUSER, D., AND SCHWARZ, N. Attentive turkers: Mturk participants perform better on online attention checks than subject pool participants. Behavior Research Methods 48 (03 2016), 400–407.

[21] HYDOCK, C. Assessing and overcoming participant dishonesty in online data collection. Behavior Research Methods 50 (12 2017).

[22] INC., A. T. Amazon mechanical turk, 10 2001. Patent Number: 7197459.

[23] IPEIROTIS, P. Demographics of mechanical turk. A Computer Scientist in a Business School [Blog] (03 2010).

[24] KHRONOS GROUP INC. Webgl. `https://www.khronos.org/webgl/`, 2011. accessed: 2020-09-03.

[25] KLEEMANN, F., VOSS, G. G., AND RIEDER, K. Un (der) paid innovators: The commercial utilization of consumer work through crowdsourcing. Science, technology & innovation studies 4, 1 (2008), 5–26.

[26] KOBLIN, A. The sheep market: Two cents worth. Final thesis for UCLA). Available online: http://www. aaronkoblin. com/work/thesheepmarket/TheSheepMarket. doc [Cited in 30.4. 2011] (2006).

[27] KUKKONEN, H., ROVAMO, J., TIIPPANA, K., AND NÄSÄNEN, R. Michelson contrast, rms contrast and energy of various spatial stimuli at threshold. Vision research 33 (08 1993), 1431–6.

[28] LANDERS, R., AND BEHREND, T. An inconvenient truth: Arbitrary distinctions between organizational, mechanical turk, and other convenience samples. Industrial and Organizational Psychology 8 (06 2015), 142–164.

[29] LANGER, M., AND BÜLTHOFF, H. A prior for global convexity in local shape from shading. Perception 30 (2001), 403–10.

[30] LITMAN, L., ROBINSON, J., AND ROSENZWEIG, C. The relationship between motivation, monetary compensation, and data quality among us- and india-based workers on mechanical turk. Behavior research methods 47 (06 2014).

[31] MACINNIS, C. C., BOSS, H. C., AND BOURDAGE, J. S. More evidence of participant misrepresentation on mturk and investigating who misrepresents. Personality and Individual Differences 152 (2020), 109603.

[32] MASON, W., AND WATTS, D. J. Financial incentives and the "performance of crowds". In Proceedings of the ACM SIGKDD Workshop on Human Computation (New York, NY, USA, 2009), HCOMP '09, Association for Computing Machinery, p. 77–85.

[33] MOSS, A., ROSENZWEIG, C., ROBINSON, J., AND L., L. Demographic stability on mechanical turk despite covid-19. Trends Cogn Sci (06 2020).

[34] OPEN SOURCE. Nodejs. `https://nodejs.org/en/download/`, 2019. version = v12.18.3.

[35] OPEN SOURCE COMMUNITY. Three js. `https://github.com/mrdoob/three.js`, 2019. commit = bcfa3339edf0222ee8b9509417c92640ce1cd3d9.

[36] OPENJS FOUNDATION. Expressjs. `https://expressjs.com/`, 2019. accessed: 2020-09-03.

[37] PELI, E. Contrast in complex images. Journal of the Optical Society of America. A, Optics and image science 7 (11 1990), 2032–40.

[38] PENTLAND, A. Linear shape from shading. Int J Comput Vision 4, 4 (1990), 153–162.

[39] RAMACHANDRAN, V. S. Perceiving shape from shading. Scientific American 259, 2 (1988), 76–83.

[40] ROBINSON, J., ROZENSWEIG, C., AND LITMAN, L. The Mechanical Turk Ecosystem. SAGE Publications, 2021, ch. 2.

[41] ROSS, J., IRANI, L., SILBERMAN, M., ZALDIVAR, A., AND TOMLINSON, B. Who are the crowdworkers? shifting demographics in mechanical turk. Conference on Human Factors in Computing Systems - Proceedings (01 2010), 2863–2872.

[42] ROSS, J., ZALDIVAR, A., IRANI, L., AND TOMLINSON, B. Who are the turkers? worker demographics in amazon mechanical turk. Department of Informatics, University of California, Irvine, USA, Tech. Rep (2009).

[43] STEWART, N., CHANDLER, J., AND PAOLACCI, G. Crowdsourcing samples in cognitive science. Trends in Cognitive Sciences 21 (08 2017).

[44] STEWART, N., UNGEMACH, C., HARRIS, A., BARTELS, D., NEWELL, B., PAOLACCI, G., AND CHANDLER, J. The average laboratory samples a population of 7,300 amazon mechanical turk workers. Judgment and decision making 10, 5 (9 2015).

[45] WAGNER. Simplex noise. `https://github.com/jwagner/simplex-noise.js`, 2019. commit: 7ec0556cc96cbc4db3f29ba0602ef0b4b9242009.

[46] WALTER, S., SEIBERT, S., GOERING, D., AND O'BOYLE, E. H. A tale of two sample sources: Do results from online panel data and conventional data converge? Journal of Business and Psychology (2018), 1–28.

[47] WOODS, A., VELASCO, C., LEVITAN, C., WAN, X., AND SPENCE, C. Conducting perception research over the internet: A tutorial review. PeerJ 3 (07 2015).

[48] ZACK, E. S., KENNEDY, J. M., AND LONG, J. S. Can nonprobability samples be used for social science research? a cautionary tale. Survey research methods 13 (2019), 215–227.

[49] ZHANG, B., AND GEARHART, S. Collecting online survey data: A comparison of data quality among a commercial panel —& mturk. Survey Practice 13 (12 2020), 1–10.