BRILL *Art & Perception (2020)* DOI: 10.1163/22134913-bja10021 brill.com/artp

Thurstonian Scaling and the Perception of Painterly Translucency

M. W. A. Wijntjes*, C. Spoiala and H. de Ridder

Perceptual Intelligence Lab, Industrial Design Engineering, Delft University of Technology, Landbergstraat 15, 2628 CE Delft, The Netherlands

Received 17 January 2020; accepted 30 June 2020

Abstract

Visual material perception is often studied with physically well-defined stimuli that lack ecological variety. Yet, even the visual variety found in our natural environment is limited when compared to artistic depiction. A similar object can be depicted in numerous different ways that all make visual sense. We studied the perception of translucency using 38 paintings of sea waves as experimental stimuli. It has previously been shown that translucency depends on the shape of the translucent object and on the light conditions. Both shape and light appear in many variations in depictions of seas. In the first experiment we explored the use of Thurstonian scaling and introduce the concept of Number of Distinguishable Levels (NDL). We found that the NDL ranged between 1.5 in a set with small waves to 4 in a set with large waves. While Experiment 1 took place in the lab, Experiment 2 was performed online and replicated the data from Experiment 1 qualitatively, although the NDL was lower in the online experiment. Furthermore, in this experiment we conducted Thurstonian scaling on a number of other attributes that possibly contribute to translucency perception, such as wavetip shading, surface reflections and realism. We found that many of these correlated significantly with translucency. In sum, this study advocates and demonstrates the use of uncontrolled stimuli, in our case paintings, to explore the wide variety of input the human visual system can process.

Keywords

Material depiction, Thurstonian scaling, marine art, translucency

* To whom correspondence should be addressed. E-mail: m.w.a.wijntjes@tudelft.nl

 DOI: 10.1163/22134913-bja10021

1. Introduction

Studies on perception are often conducted under controlled conditions. Observers are presented with physically quantified and controlled stimuli in a 'dimly lit' lab space. Relating mental responses to physically controlled stimuli is known as psychophysics, which is a standard empirical paradigm in experimental psychology. This paradigm provides answers to many questions about perception but critically relies on the imagination of the experimenter, who chooses which physical parameters to manipulate. Aside from the experimenters' imagination, the tools to generate stimuli (e.g., rendering software) also limit the wide variety of what the visual system can process.

A different approach is the use of uncontrolled stimuli. For example, 'natural images' can be used that reflect (a piece of) our natural environment. The advantage is increased variety as there is no limitation of the psychophysicist's imagination or ability (e.g., rendering expertise). Yet, our visual system is able to perceive beyond natural imagery. If we go one step further, we arrive at what could be called 'unnatural stimuli': the artistic communication of visual experience. With paintings (to name one of many media we can learn from), artists have a freedom that is not limited by physics (Cavanagh, 2005).

Although both natural images and paintings are not *a priori* controlled, it is still possible to compare their properties with visual experience. These properties can be image statistics, other behavioural data, annotations, meta data, etc. The possibilities of this approach are vast and rather unexplored. The motivation behind the current study is twofold: we aim to make progress on the methodology for this approach while at the same time gain understanding on a specific perceptual material attribute: translucency. In the following sections we will first introduce translucency perception and depiction, and secondly discuss the specific methodology we chose to explore in this study.

2. Translucency Perception and Depiction

Object appearance is determined by how light is reflected, transmitted and absorbed. Opaque materials only reflect and absorb light, while translucent and transparent materials also transmit light. The perception of both transparency and translucency have been subject to empirical investigations. Transparency was initially investigated by Metelli (1974) who could predict what combinations of lightness determine the perception of transparency at junctions. This concerns thin light-transmitting sheets (e.g., of glass or fabrics) and Sayim and Cavanagh (2011) showed that Metelli's rules were known quite well throughout art history.

In contrast with the artists' (implicit) knowledge about transparency of thin materials, Sayim and Cavanagh (2011) argued that artists seemed to ignore the

physical implications of thick materials: light refraction in volumetric materials (e.g., a water-filled glass vase) is often misrepresented. Refraction leads to a general distortion of the background image, which is one of the cues for transparency in the volumetric case (Fleming *et al.*, 2011) as opposed to Metelli's thin, flat sheets. Although the role of these distortions with respect to material estimation is debated (Schlüter and Faul, 2014), they at least contribute to 3D shape perception (Schlüter and Faul, 2019). Besides background distortions, specular reflections and light absorption also determine the appearance of transparent shapes (Schlüter and Faul, 2019).

Transparency is a special case of translucency. Transparent and specular materials reflect or transmit the whole environment and the distortions and reflections contain (albeit distorted) spatial statistics of this environment. Translucent and matte objects typically loose all higher-order spatial information of the environment [see Basri and Jacobs (2003) for the Lambertian case] and are thus shaded 'smoothly'. However, smooth shading gradients is where the comparison ends between matte and translucent objects, as their relations between shape and shading are very different (Koenderink and van Doorn, 2001; Marlow *et al.*, 2017). The shading pattern of translucent materials depends on the geometry behind the surface, while for opaque materials only the surface geometry matters. For translucent materials, light enters the subsurface and is then diffusely scattered inside the medium before either exiting the material or being absorbed. This process makes translucency depend on the thickness of shapes and thus also on the scale of objects: a small cube contains relatively much 'thin' areas compared to a large cube (Fleming and Bülthoff, 2005). The same authors also note that the sharpness of (cast) shadow edges becomes more blurry, which is interesting in the context of sea paintings. Although transparent and translucent materials have vastly different shading patterns, both the appearance of transparency (Schlüter and Faul, 2019) and translucency are affected by the presence of specular highlights. For translucent materials, the presence of highlights increases the perceived magnitude of the translucency (Motoyoshi, 2010). Furthermore, perception of translucency is critically dependent on lighting and viewpoint (Xiao *et al.*, 2014).

To investigate translucency perception using artistic depictions, we choose the topic of sea paintings. Depictions of seas can be found throughout art history, serving as a stage for explorations, battles and metaphors. The depiction of sea water involves (implicit) knowledge of both optics and fluid dynamics. Seawater can vary considerably between transparent and translucent. How relevant are the visual cues described in the perception literature for the depiction of seawater? Some images readily come to mind where reflections of a sunset sparkle over the ocean's surface. Although highlights influence perception of transparency (Schlüter and Faul, 2016) and translucency (Motoyoshi, 2010),

they only do so in combination with background distortions (transparency) or shading (translucency). Besides this complicating factor that 'cues' may not necessarily add up linearly but rather rely on specific combinations to affect translucency, we also cannot factor out the influence of semantics and/or cognition. It is quite likely that observers, when judging the translucency of seas, take into account non-visual aspects, i.e., when scene context is the Caribbean, the sea is likely more translucent than in a typical North Sea scene.

To further illustrate the connection between translucency cues described in vision science and the depiction of seas we will shortly discuss a few cases. In the experiments reported below we grouped paintings into four categories to limit some of the variability. Specifically, the shape of waves likely influences the perception of translucency. The categories were labelled Small, Shore, Large and Storm. The Small waves (three out of 10 are shown in Fig. 1) potentially contain three relevant cues for translucency. In the top painting the shadow edge in the foreground/right part is rather blurry. Vague shadow edges can have various origins, in this case it is most probably caused by the blurry edge of a cloud. Nevertheless, blurry edges are also typical for translucent materials (Fleming and Bülthoff, 2005) and we do not know how human observers weigh the potential causes of cues (if they do that at all). Secondly, the contrast of the waves (in the two different 'light zones'; Kartashova *et al.*, 2018) is relatively low. Because of the internal scattering in translucent media the contrast is indeed lower in translucent than opaque media (Motoyoshi, 2010). Thirdly, the person in the foreground can partially be seen through the water. Whether this is the refracted lower body or reflected upper body is unclear, so potentially ambiguous. Thus, for observers resolving the ambiguity by assuming refraction, this can be an important cue.

In the Shore paintings, waves break while approaching the shore. Besides contrast, shadow edge, refraction and specular reflections, these waves are often particularly shaped while breaking. The top of the wave can be 'looked through', and thus shows similar effects as cube edges discussed by Fleming and Bülthoff (2005). These edges light up when light comes from behind (Xiao *et al.*, 2014). As can be seen on the top painting in Fig. 1, this effect seems to be absent. The wave top is actually darkened as if vignetted by the rolling wave. The middle painting shows the opposite: the wave is dark at the bottom and light at the top, which is more in line with what we know about translucency. The bottom painting combines the translucent top wave with refracted subsurface rocks.

The third and fourth group are somewhat similar: Large and Storm. The difference between the sets is relatively small, and one painting of Large waves, shown at the top of Fig. 1, was actually part of both sets. As previously discussed, the shading patterns of translucent media are much more complex than those of opaque materials because 'translucent shading' depends on the

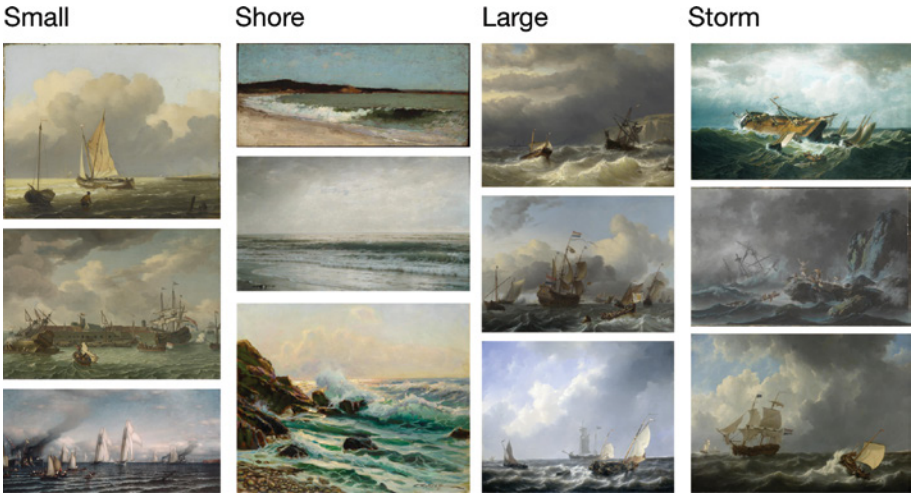

Figure 1. Examples of the 4 sets used in the experiments.

geometry behind the surface. On the other hand, this complexity vanishes when it comes to open seas as beneath the surface there is optical infinity. Therefore, rendering light wavetips might actually be the effective cue for a translucent sea. Another (complementary) shading strategy could be the use of smooth gradients in any directions as long as it does not covary deterministically with the surface attitude, as would happen to opaque matte materials.

3. Thurstonian Scaling

Thus, these four sets of paintings all contain potential translucency cues. But how to investigate the perception of these uncontrolled stimuli? We chose to use Thurstonian scaling (Thurstone, 1927) which orders stimuli on the basis of discrimination thresholds. This means that if two stimuli are indistinguishable their location on the resulting perceptual scale is close, and *vice versa*. A disadvantage is when the set of stimuli is too easy to discriminate (known as supra-threshold stimuli), the distances become ill defined. Another disadvantage of Thurstonian scaling, is that the resulting scale is not necessarily perceptually homogeneous (Knoblauch and Maloney, 2008). That discrimination thresholds reflect actual perceptual differences is an assumption that may not always be true. To overcome this problem, Maloney and Yang (2003) proposed Maximum Likelihood Difference Scaling (MLDS). Here, observers directly judge which of two stimulus pairs is more similar, therefore directly addressing perceptual differences. However, our interest is different from MLDS users. MLDS has only been used on controlled stimuli where the

purpose was to create a mapping between a physical parameter and perception. As we do not have *a priori* parameters, this homogeneity is not necessarily of our interest. Furthermore, in Experiment 2, we will use Thurstonian scaling to map the strength of available cues in the paintings, and relate these to perceived translucency. The comparison of translucency with cues achieved through the same scaling method overcomes the problem of perceptual homogenous mapping. A practical problem of MLDS is that the stimuli should be ordered *a priori*, which is typically unproblematic for controlled stimuli, but obviously impossible in our case. Therefore, we believe that Thurstonian scaling is a suitable and perhaps even an exciting method to address our questions as we will introduce three new contributions concerning this method. Two are concerned with modelling and will be further addressed in the data analysis section. One contribution is more conceptual. The 'disadvantage' (Knoblauch and Maloney, 2008) that Thurstonian scaling uses discrimination thresholds seems an advantage in our case, as the resulting scale directly tells us how many levels of distinguishability are present in a certain set of stimuli. It seems of general interest to know how many distinguishable levels are present within stimulus sets in a wide variety of contexts. For example, to quantify illuminants one may compare the number of distinguishable levels of gloss or colour under two or more different illuminants. In our context, the number of distinguishable material qualities can be compared between schools, time periods, artists, medium etc. For example, it could be that in the Renaissance only two levels of gloss were depicted within a certain material class (e.g., metal) while in the 17th century, ten different levels. Although we have not formulated specific art-historical questions, our exploration of Thurstonian scaling may facilitate future research to do so.

4. Experiment 1

4.1. Methods

4.1.1. Observers

Twelve observers (three males, nine females) participated in the experiment. They were naive with respect to the purpose of the experiment.

4.1.2. Stimuli and Apparatus

The paintings were selected by the authors, primarily on the basis of visual criteria, rather than art-historical. We browsed through museum collections that were Open Access and first selected hundreds of sea paintings. Then we defined four categories and reduced the selection to 38 paintings: (1) small waves ($n = 10$); (2) shore waves ($n = 9$); (3) large waves ($n = 10$); and (4)

stormy waves ($n = 10$). One painting was categorised in both the Large and Storm set, hence the total of 38. The rationale behind this categorisation is mostly discussed in the Introduction. It could be questioned whether the small apparent differences between the Large and Storm sets justify defining two different sets. From a practical point of view, we need to create sets of about 10 stimuli and whether the distinction makes sense is up for testing.

Images came from four collections: one painting from the Getty Museum, 12 from the Metropolitan Museum, nine from the National Gallery London and 17 from the Rijksmuseum. We found these by searching throughout the collections of these four museums. Fourteen paintings were from the 17th century, three from the 18th century, 18 from the 19th century and three from the 20th century. All details can be found in the appendix.

The images were presented pairwise on a 27-inch LCD monitor (CG277; EIZO Corporation, Hakusan, Japan) in a dimly lit room. The experiment was written in HTML, using the P5.js library (McCarthy *et al.*, 2015) for graphic presentation. Thus, the experiment was run from a webserver. The images were downscaled to 600 pixels wide while keeping the original aspect ratio. Observers could click the image of their choice upon which the next trial was presented.

4.1.3. Procedure

The experiment consisted of two parts. In the first part, observers performed the pairwise comparison task, in the second part they were interviewed and asked to reflect upon their choices. The pairwise comparison task consisted of four blocks presented in quasi-randomised order. The instructions to the participants were:

You will be presented with a series of paintings of the sea. We will ask you how well the painter depicts the translucency of the water. With translucency, we mean the opposite of opaqueness, but we are not limiting to pure transparency. For example, tea with milk is still more translucent than a cup of white paint.

Observers were presented with four repetitions of the same pair, amounting to 180 trials for the 10-stimulus sets (Small, Large and Storm), and 144 trials for the nine-stimulus set (Shore). The left–right positions of the paintings were randomised trial-by-trial. Observers clicked the image that depicted the sea that was rendered most translucent and continued automatically to the next trial.

After finishing all four sets, the experimenter interviewed the participants about their considerations when doing the task. A print of each stimulus was shown and they were asked on what they based their responses. Observers were free to talk about any facet of the depiction.

4.1.4. Data Analysis

The pairwise comparisons can be used to construct an ordering: the sea that receives the most 'votes' is the most translucent (in this case), and the one receiving least is least translucent. Thurstone (1927) introduced a way to extract more information from a pairwise comparison experiment then mere order: Thurstonian scaling allows to construct perceived 'distances' on the basis of discrimination thresholds. When we assume that the responses behave according to the Gaussian normal distribution, then the Gaussian cumulative distribution function (see Note 1) can be used to convert between response and distance.

In Fig. 2, this function is shown. On the *y*-axis are the responses, and on the *x*-axis the perceptual distance (Tsukida and Gupta, 2011). The distance unit is actually the standard deviation of the Gaussian, which means that 1 relates to the 84% level, which is called the Just Noticeable Difference (JND) (Note 2).

An example of Thurstonian scaling on our data is shown in Fig. 3. The technique of Thurstonian scaling is relatively standard but in this paper we introduce three new ideas. The first one is more conceptual and concerns the usages of Thurstonian scaling for research on art perception. The other two are more technical.

Firstly, we introduce the concept of Number of Distinguishable Levels (NDL). As Thurstonian distances are defined by the JND, i.e., the discrimination threshold), the total length of the Thurstonian scale reflects the number of JNDs, i.e., the number of distinguishable levels. As mentioned in the Introduction, this number can change over various interesting independent variables related to either the stimulus set (e.g., comparing two painters) or the specific instruction (e.g., using a different attribute).

Secondly, we use an additional step in the fitting procedure that increases the accuracy of the scale values and gives insight into individual variances of

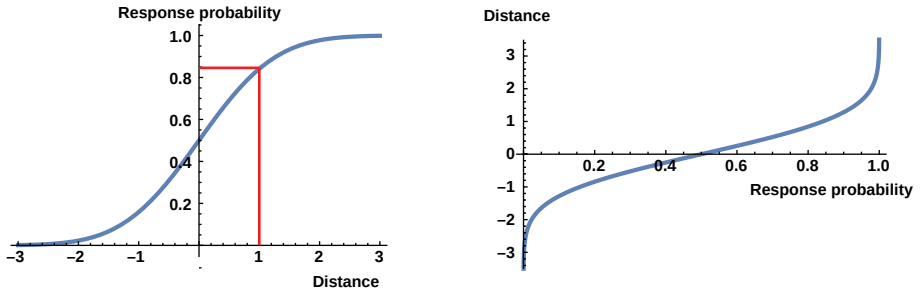

Figure 2. On the left, the cumulative Gaussian distribution which is described by two parameters: the standard deviation (σ) and bias (μ). In this case, we set $\sigma = 1$ and $\mu = 0$. The inverse of this function is used to convert pairwise comparison answers (response probability) to a distance.

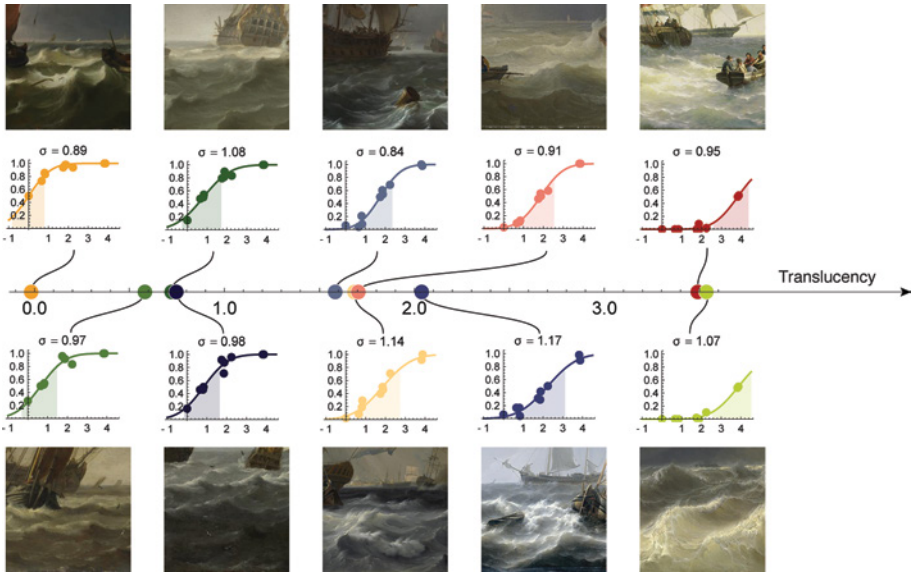

Figure 3. Visualisation of Thurstonian scaling and psychometric functions. The data comes from the 'Large wave' set, and the images are details from the paintings. In the middle, the Thurstonian scale is shown. This scale is used for psychometric fitting, thus each psychometric curve shows the result of that particular painting (with probability 0.5) in comparison to the rest. The coloured area denotes (twice) the threshold value, i.e., the range around the stimulus where comparisons range between 16% and 84% correct. It can roughly be seen that about two of these areas fit the whole range, which is thus about 4 JNDs (Just Noticeable Differences) long.

the stimuli. Normally, the so-called Case V model (Thurstone, 1927) is used, which assumes equal variances for all stimuli. As shown in Fig. 3, the data and model are actually a set psychometric functions for each stimulus. Psychometric function fitting (Wichmann and Hill, 2001a) allows to fit both the 'bias' (horizontal shift) and 'threshold' (slope of the sigmoid). We used this fitting procedure with thresholds and biases as free parameters. Off course, on average the thresholds should equal 1, as imposed by Thurstonian scaling, i.e., by using the conversion of response to distance from Fig. 2. This approach allows to see if the Case V assumptions are justified, and whether certain paintings cause more variance than others. The fitted biases should ideally be the scale values, but can differ from these for two reasons. The first is that the transformation of response to distance as shown in Fig. 2 runs into problems with responses that are either 0 or 1. As the associated distances are minus and plus infinity, the normal procedure is to cap these values by a certain limit. This is mostly solved when including all data in the fitting, as there will generally be a number of responses larger than zero and smaller than one. The second reason the biases can change is because the thresholds are now free parameters.

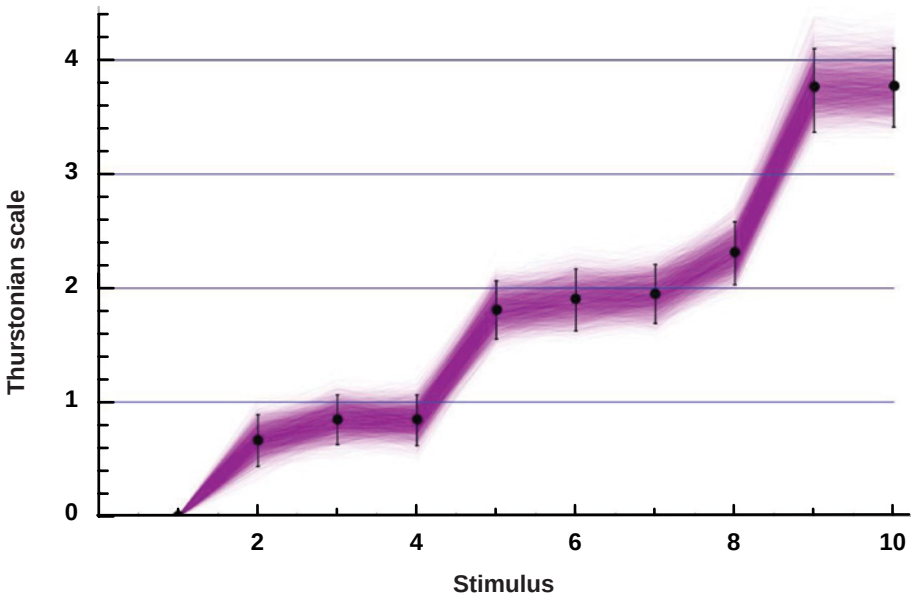

Figure 4. Visualisation of Monte Carlo simulations of Set 3. The purple lines show synthetic data generated by Monte Carlo simulations. The error bars indicate the lower and upper 95% confidence intervals, which were computed numerically (i.e., simply calculating the 2.5% edges in the data.)

The third contribution is related to the concept of NDL and the psychometric fitting we employed. We would like to know a confidence interval for the NDL which can be estimated with Monte-Carlo simulations. This seems a logical consequence of formulating Thurstonian scaling as psychometric fitting. We used the procedures described by Wichmann and Hill (2001b) and ran 2000 simulations for each set of stimuli. This is visualised in Fig. 4. Note that the 95% confidence intervals (the error bars) are substantially smaller than the JND levels, i.e., the Thurstonian scale values.

4.2. Results

4.2.1. Scaling

First we computed the NDLs of the four different sets of stimuli, shown in Fig. 5. As can be seen, the NDLs vary quite a lot, from 1.6 in case of the Small waves to 3.9 in case of the Large waves. The 95% confidence intervals indicate that all sets differ from each other except the Large and Storm sets.

Furthermore, we plotted all scale values including the individual JNDs in Fig. 6. In the Large set, there appear to be four subgroups within the set, each separated by approximately 1 JND. The other groups show more of a continuum. Furthermore, it can be seen that the error bars vary substantially, which

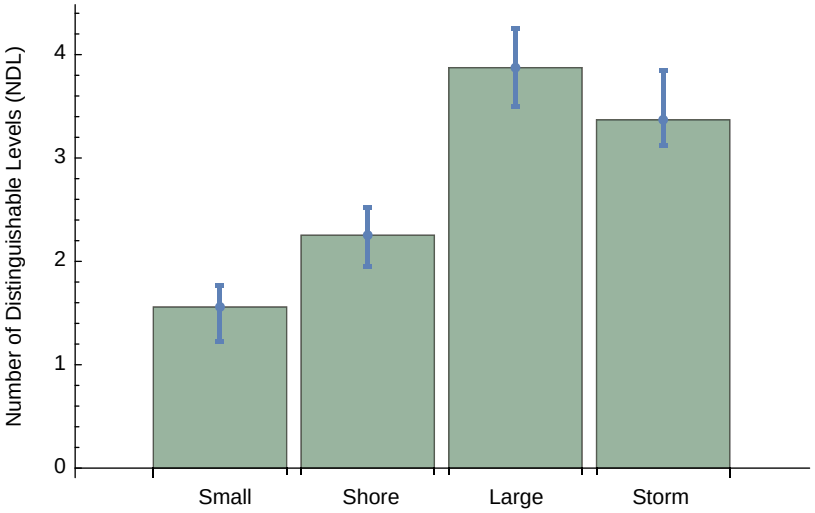

Figure 5. Number of Distinguishable Levels (NDLs) for the four sets of paintings. The set of paintings with small waves had fewest distinguishable levels of translucency while the large waves contained more than twice as many. The error bars indicating the 95% CI are based on Monte Carlo simulations.

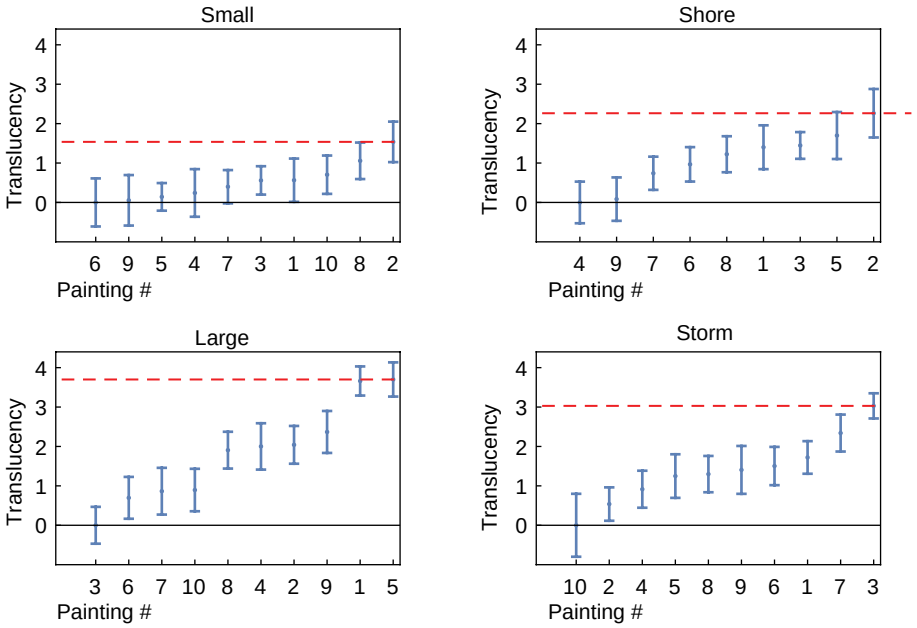

Figure 6. Levels of perceived translucency in terms of JND. Error bars denote individual JNDs per picture (and not the 95% CI). On the *x*-axis the painting labels are shown, and ordering is with respect to translucency. The red dashed line indicates the maximum value, thus denoting the number of discriminable levels.

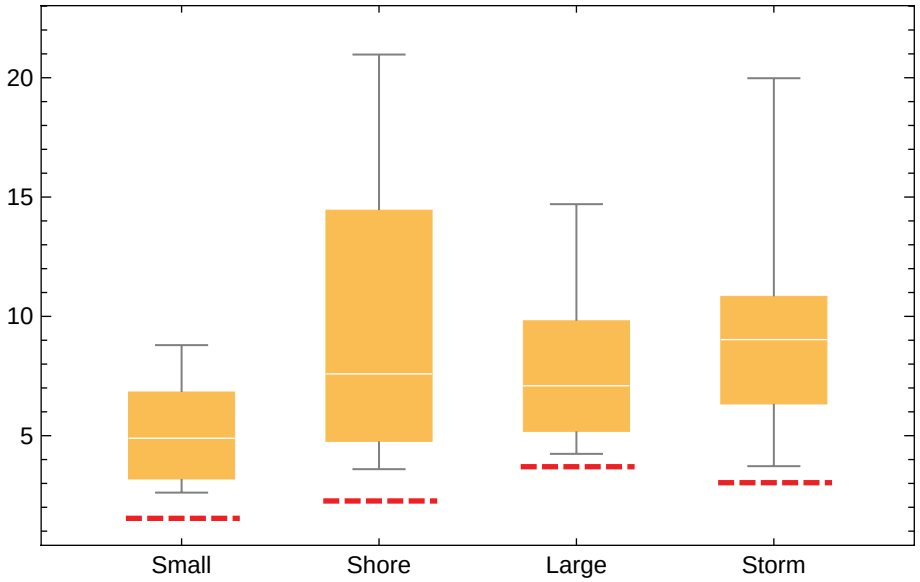

Figure 7. The Number of Distinguishable Levels for the group data (red dashed line) and individual data shown in box–whisker plots, where boxes represent the 25% and 75% quantile and the white line denotes the median.

seems to suggest that the Thurstonian Case V assumption of equal variance does not hold.

Next, we were interested whether the Number of Distinguishable Levels based on the aggregated group data would be similar to the NDLs based on individual data. To do so, we computed the levels for each observer in each condition. The box–whisker plot of this analysis is shown in Fig. 7. As can be seen, all individual NDLs are higher than group data (red dashed line). Note that these numbers in some cases exceed the number of stimuli, suggesting that all stimuli can be distinguished above the 84% level. Values larger than the number of stimuli may not be the best estimate of the 'true' value as fitting psychometric curves to data that are almost always correct can be unstable. Another interesting aspect is that the NDL for the group data (red dashed line) is highest in the Large set, while the NDLs for the individual data (white line in box–whisker plot) is highest in the Storm set.

4.3. Introspection Analysis

After the experiments, observers were interviewed while presented with all images and asked to reflect upon their responses. First, we tried to extract the level of perceived translucency from the transcribed thoughts. Observers often used words like 'highest', 'high', 'medium', 'low', 'lowest' which was

transformed to a five-point scale. We collected these per image and computed the correlation between these introspective data and the Thurstonian scale values. Perhaps surprisingly, this correlation was highest for the Small-wave set ($r = 0.91$) while lowest for the Large-wave set ($r = 0.58$) and intermediate for the Shore ($r = 0.77$) and Storm ($r = 0.73$) set. Next, we studied their introspections in more detail and found six topics that were repeatedly mentioned:

1. **Realism**: Seas that received a low translucency score were also perceived as unrealistically painted. In some cases, the water was perceived as a different material (e.g., rock, stone, plastic, sand, steel, or petrol). When translucency was strong, observers also remarked that the depiction was realistic.
2. **Light**: Some observers noted that the weather conditions made translucency come out better, i.e., when the sun was shining. Also, certain light spots (possibly sunlit parts surrounded by cloud shadows) were mentioned to contribute to translucency. Furthermore, a shadow was remarked as increasing translucency.
3. **See-through**: Actually seeing something through the water was discussed as an important ingredient for translucency. Sand, beach, rocks, boats and persons were (partly) seen through the water.
4. **Shading:** In many cases, observers noted that the top of the waves of translucent seas were shaded in a light tone, showing the light passing (and scattering) through a 'thin' volume. Also, observers made notice of a certain 'glow' of the seas that were perceived as translucent.
5. **Reflections:** Highlights were noted but their role not always interpreted similarly among observers. One observer said that because of the many reflections, the sea was judged opaque. Another observer was more ambivalent, saying that it could simply not be said how translucent water is that is covered by highlights, noting it might as well be 'black [opaque] liquid'. Still others deemed highlights positively diagnostic for translucency.
6. **Contrast:** Observers mentioned contrast, mostly that increased contrast indicated more translucency.

4.4. Discussion

In the first experiment, observers viewed 'uncontrolled' stimuli under controlled lab settings. The pairwise comparisons resulted in translucency scales for the four sets of sea paintings. We introduced the concept of NDLs. While this number can also depend on the attribute/instruction, we only varied the stimulus set, which revealed a relatively large difference between sets: the smallest and highest NDLs differed by a factor of 2. The difference between the sets is largely the shape and size of the waves, at least that was the criterion we used when we selected the paintings. It would perhaps be better if we could

in some way quantify these differences, for example by letting observers rate the shape/size of waves. For now, we trust our own classification judgements; the reader is free to verify this in the supplementary material.

Assuming that the shape of waves indeed varies across the four sets, then the different NDLs are likely due to these differences in waves. Apparently, large waves allow for more levels of translucency than small waves. The shapes of large waves seem to allow more levels of translucency to be depicted and perceived. This is interesting beyond the current scope of translucency perception as many other questions can be answered using this paradigm. For example we could also show a number of isolated patches from a single sea and perform Thurstonian scaling. If the NDL is large, then translucency perception substantially depends on local information present on certain patches.

Furthermore, we showed that individually, the NDLs can be substantially higher than for group data. This shows that judgements are partly idiosyncratic. The relative increase for the individual JND lengths is only a factor 2 for the Large-wave set, and around a factor 3 to 4 for the other sets. Thus, the Large-wave set seems least ambiguous.

Lastly, we asked observers to introspect on their responses and contemplate what kind of information they used to make their translucency assessments. This resulted in interesting insights that partly overlap with existing literature on translucency perceptions, such as the light shading at thin areas (the wavetips) and the contribution of highlights.

While these findings seem quite interesting to us, they also generate a number of questions that we would still like to address within the scope of the current study. Primarily, we wanted to test whether the qualitative insights are also quantitatively related to translucency judgements. At the same time, we were also interested in running the experiment outside the controlled lab environment, as we now needed a substantial amount of data. Therefore, we decided to replicate the data from two of the four sets of paintings and to additionally run Thurstonian scaling experiments on five potential factors that may contribute to translucency, as found in the qualitative data.

5. Experiment 2

5.1. Methods

5.1.1. Participants

We used Amazon Mechanical Turk (MTurk) to recruit participants for our study. This is an online platform for crowdsource work, used for so-called "Human Intelligence Tasks", known as a HIT. This HIT can be a task to label traffic signs in videos, translate audio or transcribe scanned handwritten documents, but is

also used in a variety of psychological studies. In our case, the HIT is an experimental block in which we ran the paired comparison experiment. The reward for this HIT was set to $1. In total we recruited 20 participants for our 12 HITs (see below for explanation), totalling to 240. The total number of unique participants was 207 which means that some observers participated in more than one HIT. Nineteen of the 207 observers performed two HITs, six performed three HITs and one performed four HITs.

We do not have information about the participants' gender and age. We did use three criteria to select participants through the MTurk platform: (1) US-based, (2) finished 1000 HITs successfully and (3) had an acceptance rate of 95%. These are relatively standard criteria that both ensure the observers understand the language and a certain level of dedication to perform the experiment seriously.

5.1.2. Stimuli and Procedure

As described in the previous section, we conducted all experiments via MTurk. This implies that observers used various types of screen sizes, resolution and viewing distances. Furthermore, the physical environment can vary, and we cannot exclude the possibility that participants were focussing on more activities than our experiment. Aside from these factors, the experiments were similar. We used the same HTML web page to conduct the experiments, and the stimuli of Set 1 and Set 3 from Experiment 1. Aside from repeating these sets from Experiment 1, we conducted pairwise comparison experiments on five qualities that could be related to the translucency judgements. The instructions for these experiments are:

1. **Realism:** We will ask you how convincing the painter depicts the sea. With convincing, we mean something like realistic, but it does not necessarily have to be 'real', i.e., a dragon can be painted very convincing but a dragon is not real.
2. **Wavetips:** We are interested in a particular aspect of how the sea has been painted: the tip of the waves. The wavetips are sometimes painted lighter than its surrounding (for example because light passes through). A cartoon of this effect is shown below.
3. **Light–shadow:** We are interested in a particular aspect, which is the contrast between sunlit and shadowed areas. In many sea paintings, the shadows of the clouds are visible on the water surface. Through this [experiment], we would like to perceptually quantify the contrast between light and shadow.
4. **Reflections:** For this HIT, we are interested in the reflections on the water surface. Reflections can be highlights caused by direct sunlight, but also mirror images of things that are reflected on the water surface.

5. **See-through:** For this HIT, we are interested in whether you can discern something underwater (e.g., a fish, piece of boat, stone, etc). It is relatively difficult as you have to compare two paintings, and decide in which painting you discern something underwater the clearest.

5.2. Results

Before we performed the actual data analysis we looked at the timing of the participants. We collected timestamps of responses and could calculate various temporal statistics. We choose to analyse the fasted 10% of the trials, with the goal of filtering out observers who were not participating seriously on the basis of these time data. Most of the response times were around 1 s but some observers were suspiciously fast. We set the threshold at 400 ms and removed nine datasets from the data (from the total of 240).

5.2.1. Number of Distinguishable Levels

The main results are shown in Fig. 8, including the data of Experiment 1. As can be seen, the results for the translucency experiment are in line with

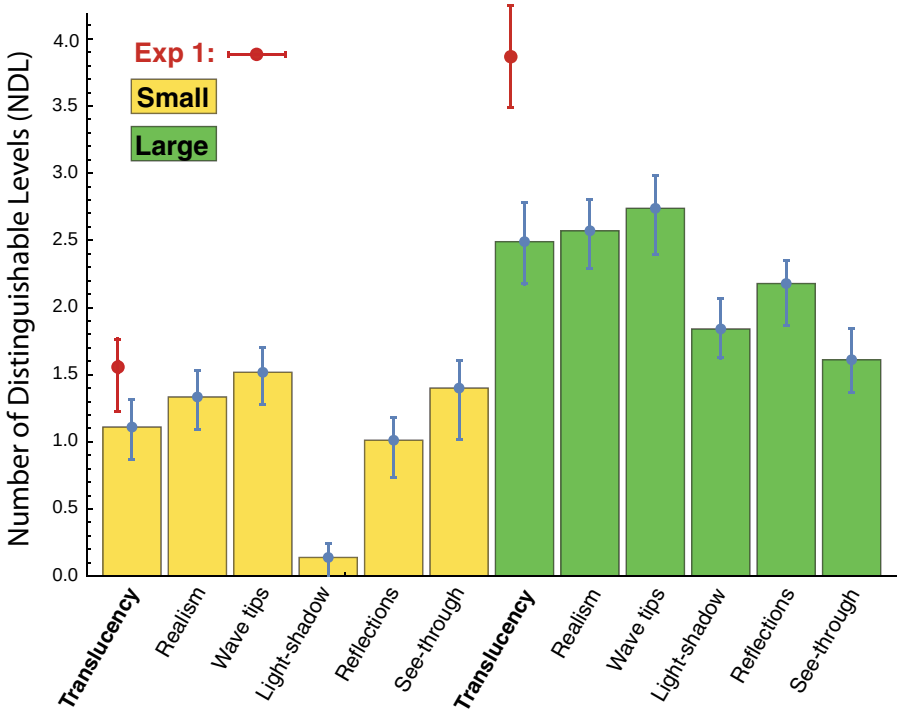

Figure 8. NDL results of Experiment 2. On the *x*-axis the attributes are shown, with in bold typeface translucency, indicating the main experiment. The red dots and errors bars denote results from Experiment 1.

Experiment 1, although the NDL is lower in the online experiment. Furthermore, the NDL of the other attributes were of the same size as translucency, around 1 for the Small set and 2 in the Large set.

5.2.2. Correlations

For each painting in each experiment we computed the 95% confidence intervals of the scale values. We correlated to translucency scales for Experiment 1 and 2 (seen in Fig. 9). For the Large waves we have near perfect correlation and for the Small waves it is slightly less.

Next, we correlated the translucency scales with the five attribute scale values (seen in Fig. 10). Furthermore, a linear model fit was performed to see how well translucency can be modelled on the basis of the visual attributes (Model 2). The See-through cue correlated rather well in both sets, indicating that this was a very important cue. As we were interested how well the model would perform without the see-through attribute, we also performed modelling without this attribute (Model 1).

As can be seen, in the Small set the correlations are lower and the Light–shadow and Reflections cues do not correlate. In contrast, the Reflections cue

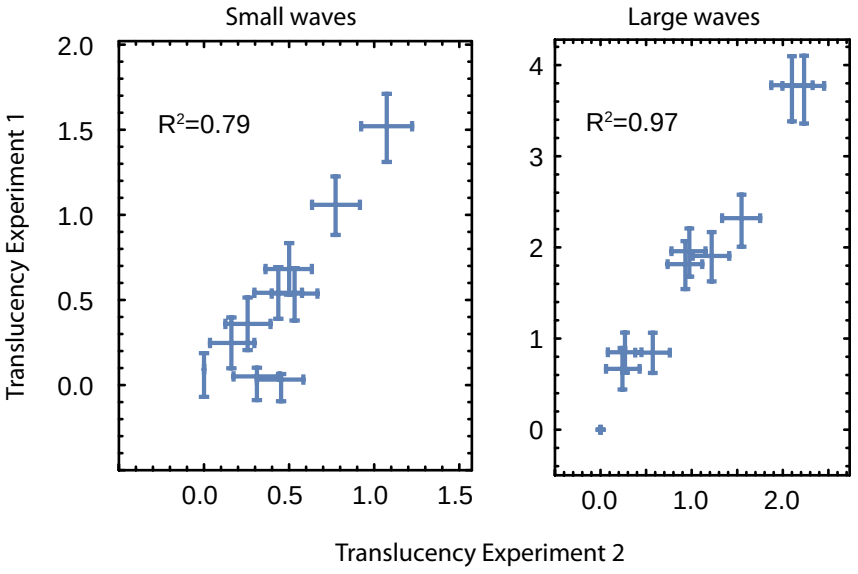

Figure 9. Correlations between data from Experiments 1 and 2. The error bars of the translucency scale values are 95% CI resulting from the Monte Carlo simulations. Mind that the aspect ratios in the plots are according to scale values, but that the range is different between the Small and Large waves.

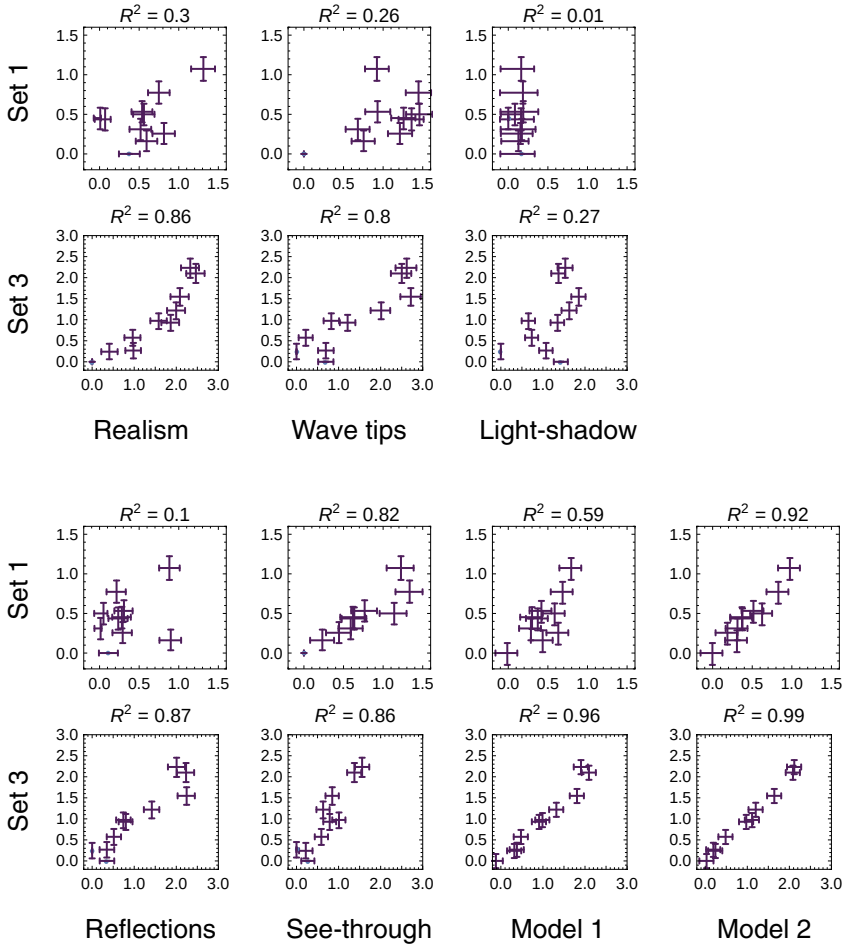

Figure 10. Correlations of the Thurstonian scales between cues (on the *x*-axis) and translucency (*y*-axis). Model 1 excluded the see-through cue, while Model 2 included all cues.

correlates as strong as most other cues in the Large-wave set. Furthermore, we found that in the Small-wave set, the model fit increases substantially when including See-through, while the increase was smaller in the Large-wave set.

5.3. Discussion

Especially for the Large-wave set, we found a substantial difference in NDLs between the lab experiment (Experiment 1) and the online experiment (Experiment 2). It is to be expected that in a controlled environment, observers are better concentrated because there may be less distractions. Also, observers

may simply take the task more seriously as they meet the experimenter in person and understand it is a serious experiment. We believe these kinds of factors may drive the difference in NDLs between the two experiments. On the other hand, it is difficult to characterise the generic online participant. In our experience, including other studies, a substantial number of participants ask attentive questions and give enthusiastic comments about doing art and perception experiments.

Although smaller, the NDLs are certainly above chance and the relative difference between the Small- and Large-wave set is also similar. Furthermore, we found high correlations between the first and second experiment, implying the observers from two totally different pools (students at TU Delft, The Netherlands, and USA-based MTurk users) behave very similar.

We continue the discussion about cue correlations in the general discussion.

6. General Discussion

We studied the perception of translucency by using various sea paintings. Sea water is both physically and pictorially a rather interesting subject as the same material exhibits different optical and geometrical properties. The shape of waves is a classic subject (Longuet-Higgins and Cokelet, 1976) and the range on both temporal (from tidal to seconds) and spatial (from cm to km) scales is very large (Munk, 1950). To minimise some of these variations, we defined four wave categories that could facilitate data interpretation. We collected both quantitative and qualitative data yielding insights in how many levels of translucency could be seen and what kind of cues observers took into account.

One of the reasons behind combining art and perception is that artists stretch the limits of what is visually possible and therefore teach us something about perception. Another reason for combining art and perception is that the perceptual results may contribute to the analysis of the artwork. The current study primarily focussed on studying perception and mostly neglected the art-historical part. The reason behind this is that discussing the methodology of Thurstonian scaling, NDLs and translucency perception already became quite complex. Yet, our results can easily serve as a starting point for a study that is more art-historically motivated.

Pairwise comparisons for picture perception have been used in other fields, such as image quality (e.g., Mantiuk *et al.*, 2012), but are not so often used in studies on material perception. Instead, MLDS (Maloney and Yang, 2003) is often preferred (Aguilar *et al.*, 2017; Fleming *et al.*, 2011; Obein *et al.*, 2004). As explained in the Introduction, there are a number of advantages of MLDS over Thurstonian scaling. Aside from possible problems with supra-threshold stimuli, we believe the difference in what both methods actually measure is important to briefly discuss first here. Thurstonian scaling uses JNDs as a

metric for the scale, while MLDS uses perceived differences in appearance. An ideal psychophysical method indeed would map a parametric scale to an appearance scale and using Thurstonian scaling in that scenario may not be appropriate. Yet, in our context of uncontrolled images the exact structure of the appearance dimension is of less interest. In fact, we believe that the NDL, which automatically follow from Thurstonian scaling, is a useful statistic to characterise sets of images. Especially in combination with confidence interval estimation we adapted from Wichmann and Hill (2001b), the NDLs can characterise differences between a wide variety of sets. For example, within a bunch of grapes (e.g., Di Cicco *et al.*, 2019), the NDL of translucency or glossiness could be measured and related to the artist. If the NDLs are high, this could indicate that the painter puts effort in rendering a variety of different grapes. But also for controlled stimuli NDLs could be useful, for example in quantifying the ability of various screens in rendering material qualities. A high-contrast, high-dynamic-range and wide-colour-gamut screen very likely can depict more distinguishable levels of materials.

The advantage of MLDS is that it does not suffer from supra-threshold pairs because it compares the relative difference between pairs. When all stimuli are supra-threshold, all pairwise comparisons will receive 0 or 1 values which makes it impossible to fit. As the studies using MLDS have used rendered stimuli where only one parameter is manipulated, it is indeed possible that observers may notice all difference accurately. Yet, in 'uncontrolled' stimuli such as our depicted sea waves, we did not expect this to happen. Indeed, on a group level we did not find supra-thresholds to influence the scaling. Yet, on an individual level, as the NDL was sometimes rather high, supra-threshold behaviour may have occurred. When a set of 10 images results in an NDL of nine, for example, then on average each neighbouring stimulus pair receives a probability of 84.1%. Every neighbour-of-neighbour pair should be chosen correctly 97.7% of the time. Taking into account that for the individual data only four repetitions were used, ceiling effects because of supra-threshold distances have likely affected the data. The high number of distinguishable translucency levels in Fig. 7 should therefore be interpreted with caution. What can safely be inferred from this figure is that the average scaling results are all substantially lower than individual data. The reason is probably due to different use of cues and their weights when making translucency inferences.

In Experiment 2 we mostly confirmed the introspections found in Experiment 1. Perhaps surprising is the link with perceived realism. This relation obviously never surfaces when using photographed stimuli, as they are all 'real'. In certain cases, observers were able to express that a different material was perceived, thus ignoring semantic influences as they are all clearly seas. Whether the absence of translucency affected their material (mis)classification

or *vice versa* cannot be inferred; both scenarios seem plausible and could also work in tandem.

The remarks about the sun and clouds is somewhat in line with our hypothesis: blurry shadow edges may be diagnostic for translucency (Fleming and Bülthoff, 2005). It is not certain that observers actually referred to this phenomenon, but if so, this is an interesting case. As discussed in the Introduction, the blurry shadow edges originate from the fuzzy cloud borders and would thus also be blurry on an opaque surface. However, Experiment 2 did not show much correlations of translucency. This can have various reasons, including our instruction which focussed more on the contrast between light and dark areas than the blurriness of the transition between these areas.

Being able to actually see an object behind the surface (which we called transparency here) is obviously strong visual evidence for something being translucent. Although perhaps trivial, this has not been discussed nor investigated in the literature. The role of deformations in transparent media has been studied (Fleming *et al.*, 2011; Schlüter and Faul, 2016, 2019) but what happens for translucent media seems unknown. The See-through cue had the highest correlations in both wave sets and was especially strong in the Small-wave set.

The light shading of wavetips was expected and mentioned quite often by participants. This fits well with other findings in translucency perception (Gkioulekas *et al.*, 2015; Xiao *et al.*, 2014). Nevertheless, this painterly procedure comes in different flavours, depending on the wave size and shape. Experiment 2 confirmed that the translucency and the lightness of wavetips are indeed correlated. This attribute is particularly interesting in comparison with photos of sea waves. The depicted shore waves of Experiment 1 seem also to exhibit the light-wavetip effect, which seems to look rather similar on photos of breaking waves. Yet, the waves seen in the Large-wave set are not so easily found on photos. Although this certainly requires more evidence, it seems an interesting hypothesis that the light wavetips occur more in paintings than in real life.

The reflections, or highlights, were only sometimes mentioned and it did not seem very likely that these are important. However, Experiment 2 revealed strong correlations between reflections and translucency in the Large-wave set. Apparently there is good agreement among observers as the NDL is about 2.2. Specular highlights occur when the surface of a solid or liquid is microscopically smooth. For liquids this seems always to be the case except when a thin layer of foam or dust is present. In that case, indeed, the liquid may also be less translucent as light cannot pass that layer. Thus, our results suggest that the findings of Motoyoshi (2010) concerning translucent solids may be generalised to liquids.

We explored how to study perception through uncontrolled images in the context of material perception and artistic depiction. Thurstonian scaling turned out to be a useful method to approach this challenge. However, this study also raises a number of questions. One of our motivations is to reveal perceptual insights from artistic practice. But what level of explanation (Koenderink, 2002) are we aiming at? We found that seas with large waves exhibit more 'shades' of translucency (higher NDL). Our interpretation is that larger waves afford the presence of more or stronger visual cues. Visual processing is often seen as recovering distal properties (the external 3D world) from proximal (retinal) signals (e.g., Anderson, 2011; Perdreau and Cavanagh, 2013). From a causal viewpoint, the visual cues itself should be proximal, leading to percepts and actions in the distal realm. Yet, when we let observers judge the wavetip lightness in Experiment 2, for example, this involves distal references (e.g., the shape of the wave). So are cues proximal or distal?

We argue they can be both and involve different levels of explanation. Image cues are clearly proximal properties, i.e., a statistic like skewness (Motoyoshi *et al.*, 2007). This makes them easy to quantify (e.g., the size or contrast) yet not always easy to interpret (Anderson and Kim, 2009) because they are not directly related to the 3D realm of perceptual awareness. Distal properties include both material properties like translucency or glossiness and to a certain degree also their responsible cues: although the luminance values and gradients of light (translucent) edges and bright (glossy) highlights are quantifiable, they are meaningless without information about the 3D shape they are positioned on (Kim *et al.*, 2011). Visual cues can thus be explained on different levels: proximal and distal. Marlow *et al.* (2012) asked observers to judge both glossiness and its hypothesised cues (coverage, disparity, contrast and sharpness). Although the authors argued that their cues were proximal, we believe they are distal: a coverage estimate of how large the highlight area is likely involves 3D shape estimation, for example. Distal cues are part of the 3D world and (hence) interact with each other, unlike proximal cues. Distal and proximal cues are surely related. For example, distal contrast (how observers judge contrast) and the proximal contrast (e.g., Michelson contrast) are very likely related through some (to be measured) psychophysical function. Yet it is not always straightforward what proximal information is relevant to the visual system.

The visual artist works on both the distal and proximal level. The proximal cues are painted directly on the retina, while these follow from distal observations. That painters connect these two levels is the main reason why art is so important to the study of perception. In the current study, we stayed on the distal level of explanation, showing that levels of translucency vary with shape and are related to various (distal) cues, and thus contribute to the bigger picture where proximal and distal information is integrated.

Notes

1. Another sigmoidal function probably works equally well, for example the Bradley–Terry model.
2. Mind the threshold level (in our case 84%) is not a standard, although widely used. The 75% levels is also often used. It is all relatively arbitrary as long as the level is explicitly stated.

References

Aguilar, G., Wichmann, F. A. and Maertens, M. (2017). Comparing sensitivity estimates 28 from MLDS and forced-choice methods in a slant-from-texture experiment, *J. Vis.* **17**, 37. doi: 10.1167/17.1.37.

Anderson, B. L. (2011). Visual perception of materials and surfaces, *Curr. Biol.* **21**, R978–R983. doi: 10.1016/j.cub.2011.11.022.

Anderson, B. L. and Kim, J. (2009). Image statistics do not explain the perception of gloss and lightness, *J. Vis.* **9**, 10. doi: 10.1167/9.11.10.

Basri, R. and Jacobs, D. W. (2003). Lambertian reflectance and linear subspaces, *IEEE Trans Pattern Anal Mach Intell.* **25**, 218–233. doi: 10.1109/TPAMI.2003.1177153.

Cavanagh, P. (2005). The artist as neuroscientist. *Nature* **434**, 301–307. doi: 10.1038/434301a.

Di Cicco, F., Wijntjes, M. W. A. and Pont, S. C. (2019). Understanding gloss perception through the lens of art: Combining perception, image analysis, and painting recipes of 17th century painted grapes. *J. Vis.* **19**, 7. doi: 10.1167/19.3.7.

Fleming, R. W. and Bülthoff, H. H. (2005). Low-level image cues in the perception of translucent materials. *ACM Trans. Appl. Percept.* **2**, 346–382. doi: 10.1145/1077399.1077409.

Fleming, R. W., Jäkel, F. and Maloney, L. T. (2011). Visual perception of thick transparent materials, *Psychol. Sci.* **22**, 812–820. doi: 10.1177/0956797611408734.

Gkioulekas, I., Walter, B., Adelson, E. H., Bala, K. and Zickler, T. (2015). On the appearance of translucent edges, *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Rec*. 5528–5536. doi: 10.1109/CVPR.2015.7299192.

Kartashova, T., de Ridder, H., te Pas, S. F. and Pont, S. C. (2018). Visual light zones, *i-Perception* **9**. doi: 10.1177/2041669518781381.

Kim, J., Marlow, P. and Anderson, B. L. (2011). The perception of gloss depends on highlight congruence with surface shading, *J. Vis.* **11**, 4. doi: 10.1167/11.9.4.

Knoblauch, K. and Maloney, L. T. (2008). MLDS: Maximum Likelihood Difference Scaling in R, *J. Stat. Softw.* **25**, 1–26.

Koenderink, J. J. (2002). Levels of explanation, *Perception* **31**, 1033–1036. doi: 10.1068/p3109ed.

Koenderink, J. J. and van Doorn, A. J. (2001). Shading in the case of translucent objects, Human Vision and Electronic Imaging VI, **4299**, 312–320. doi: 10.1117/12.429502.

Longuet-Higgins, M. S. and Cokelet, E. D. (1976). The deformation of steep surface waves on water. I. A numerical method of computation. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **350**, 1–26. doi: 10.1098/rspa.1976.0092.

Maloney, L. T. and Yang, J. N. (2003). Maximum likelihood difference scaling, *J. Vis*. **3**, 573–585. doi: 10:1167/3.8.5.

Mantiuk, R. K., Tomaszewska, A. and Mantiuk, R. (2012). Comparison of four subjective methods for image quality assessment, *Computer Graph. Forum*, **31**, 2478–2491. doi: 10.1111/j.1467-8659.2012.03188.x.

Marlow, P. J., Kim, J. and Anderson, B. L. (2012). The perception and misperception of specular surface reflectance, *Curr. Biol.* **22**, 1909–1913. doi: 10.1016/j.cub.2012.08.009.

Marlow, P. J., Kim, J. and Anderson, B. L. (2017). Perception and misperception of surface opacity. *Proc. Natl Acad. Sci. USA* **114**, 13840–13845. doi: 10.1073/pnas.1711416115.

McCarthy, L., Reas, C. and Fry, B. (2015). *Make: Getting Started with P5. js: Making Interactive Graphics in JavaScript and Processing*. Maker Media, Inc., San Francisco, CA, USA.

Metelli, F. (1974). The perception of transparency, *Sci. Am.* **230**, 90–99.

Motoyoshi, I. (2010). Highlight–shading relationship as a cue for the perception of translucent and transparent materials, *J. Vis.* **10**, 6. doi: 10.1167/10.9.6.

Motoyoshi, I., Nishida, S., Sharan, L. and Adelson, E. H. (2007). Image statistics and the perception of surface qualities, *Nature* **447**, 206–209. https://doi.org/10.1038/nature05724.

Munk, W. H. (1950). Origin and generation of waves, in: *Proc. 1st Conf. Coastal Eng.* Long Beach, CA, USA. pp. 1–4. doi: 10.9753/icce.v1.1.

Obein, G., Knoblauch, K. and Viéot, F. (2004). Difference scaling of gloss: Nonlinearity, binocularity, and constancy, *J. Vis.* **4**, 4. doi: 10.1167/4.9.4.

Perdreau, F. and Cavanagh, P. (2013). Is artists' perception more veridical? *Front. Neurosci.* **7**, 6. doi: 10.3389/fnins.2013.00006.

Sayim, B. and Cavanagh, P. (2011). The art of transparency, *i-Perception* **2**, 679–696. doi: 10.1068/i0459aap.

Schlüter, N. and Faul, F. (2014). Are optical distortions used as a cue for material properties of thick transparent objects? *J. Vis.* **14**, 2 doi: 10.1167/14.14.2.

Schlüter, N. and Faul, F. (2016). Matching the material of transparent objects: the role of background distortions, *i-Perception* **7**, 5. doi: 10.1177/2041669516669616.

Schlüter, N. and Faul, F. (2019). Visual shape perception in the case of transparent objects, *J. Vis.* **19**, 24. doi: 10.1167/19.4.24.

Thurstone, L. L. (1927). A law of comparative judgment, *Psychol. Rev.* **34**, 273–286. doi: 10.1037/h0070288.

Tsukida, K. and Gupta, M. R. (2011). How to analyze paired comparison data. UW Technical Report UWEETR-2011-0004, University of Washington, Seattle, WA, USA.

Wichmann, F. A. and Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit, *Percept. Psychophys.* **63**, 1293–1313. doi: 10.3758/BF03194544.

Wichmann, F. A. and Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling, *Percept. Psychophys.* **63**, 1314–1329. doi: 10.3758/BF03194545.

Xiao, B., Walter, B., Gkioulekas, I., Zickler, T., Adelson, E. and Bala, K. (2014). Looking against the light: How perception of translucency depends on lighting direction, *J. Vis.* **14**, 17. doi: 10.1167/14.3.17.