Evaluating Realism in Example-Based Terrain Synthesis

JOSHUA J. SCOTT, Victoria University of Wellington, New Zealand
NEIL A. DODGSON, Victoria University of Wellington, New Zealand

We report two studies that investigate the use of subjective believability in the assessment of objective realism of terrain. The first demonstrates that there is a clear subjective feature bias that depends on the types of terrain being evaluated: our participants found certain natural terrains to be more believable than others. This confounding factor means that any comparison experiment must not ask participants to compare terrains with different types of feature. Our second experiment assesses four methods of example-based terrain synthesis, comparing them against each other and against real terrain. Our results show that, while all tested methods can produce terrain that is indistinguishable from reality, all also can produce poor terrain; that there is no one method that is consistently better than the others; and that those who have professional expertise in geology, cartography or image analysis are better able to distinguish real terrain from synthesised terrain than the general population but those who have professional expertise in the visual arts are not.

CCS Concepts: • **Computing methodologies** → **Perception**; *Shape analysis*; Physical simulation; • **Human-centered computing** → **User studies**; • **Applied computing** → *Media arts*.

Additional Key Words and Phrases: terrain, example-based, evaluation, believability, realism

ACM Reference Format:
Joshua J. Scott and Neil A. Dodgson. 2022. Evaluating Realism in Example-Based Terrain Synthesis. *ACM Trans. Appl. Percept.* 1, 1, Article 1 (April 2022), 18 pages. https://doi.org/10.1145/3531526

1 INTRODUCTION

The creation of realistic artificial terrain has been a challenge for computer graphics for decades [34]. Real terrains are formed by complex physical processes over millions of years and it is a substantial challenge to generate plausible virtual terrain. Virtual landscapes that do not exhibit plausible features, or do not adhere to artistic direction, are unfit for practical applications, such as gaming and film. Realism in terrain generally refers to the terrain's adherence to the physical rules of the world and is an important aspect of terrain synthesis.

Evaluating the realism of terrain is non-trivial. In other work [50], we developed a new example-based terrain synthesis method based on texture optimization [30]. To this we added physically-based modelling, based on pit removal algorithms from GIS [51], with the intention of making drainage patterns more realistic. We wished to compare our new method against previous example-based terrain synthesis. As comparitors, we used the best existing methods for patch-based [54] and pixel-based [21] terrain synthesis. In order to independently test the pit removal addition, we developed a fourth method in which the pit-removal algorithm was added to to the pixel-based method [49, 50].

When assessing terrain, we use subjective believability as a proxy for objective realism. That is, we can measure only a participant's subjective belief in the realism of a particular terrain rather than an objective measure of the

Authors' addresses: Joshua J. Scott, jjscott.nz@gmail.com, Victoria University of Wellington, New Zealand; Neil A. Dodgson, neil.dodgson@vuw.ac.nz, Victoria University of Wellington, School of Engineering and Computer Science, PO Box 600, Wellington, 6140, New Zealand.

1544-3558/2022/4-ART1 $15.00
https://doi.org/10.1145/3531526

actual realism. For an experiment to evaluate realism through believability, the experiment must be constructed such that, for any pair of terrains being compared, if one is as more realistic than the other, then the same is true for believability.

We were aware of the potential of a 'feature bias' in evaluating terrain, where terrains with different features have different levels of believability, depending on the observers bias (e.g., consider the difference between the volcano Mount Fuji and the towering sandstone pillars in Zhangjiajie National Forest Park). Before running our main experiment, we undertook a smaller study to investigate whether there is such a bias and, if so, how this should influence the design of our main experiment.

We report on our two experiments:

(1) an experiment that shows that there is a feature bias in subjective evaluation of terrain which means that any comparisons must be between terrains synthesised to represent the same or the same type of terrain (Section 3);
(2) an experiment assessing the realism of various methods of example-based terrain synthesis that demonstrates that all methods can be indistinguishable from reality in some circumstances but that all can produce relatively poor results in others (Section 4).

We set the scene by summarising past work in terrain synthesis (Section 2) and discuss the challenge of assessing *objective realism* through experiments that ask questions about *subjective believability*.

The principal conclusions of our second experiment are that:

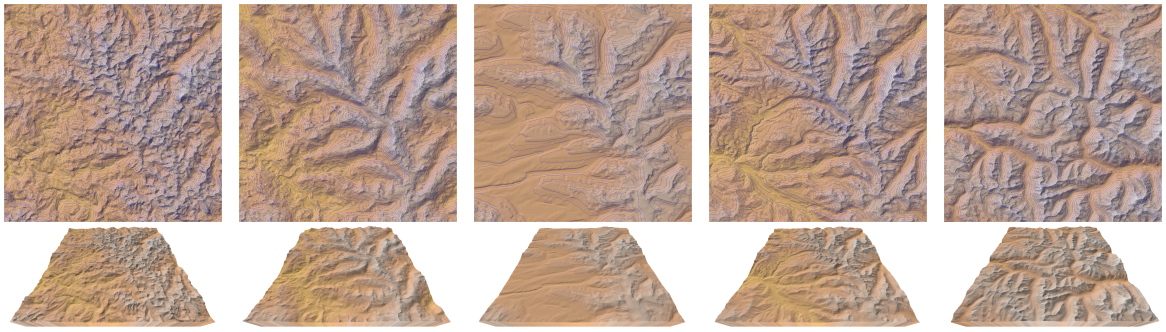
- all tested methods can produce terrain that is indistinguishable (in a statistical sense) from real terrain;
- no method can consistently achieve this level of realism;
- no tested method can be claimed to be better than any other method because all perform better than other methods on some terrains and worse on others;
- participants with professional expertise in geomorphology, cartography, or image analysis are better at detecting real terrain than other participants but participants with professional expertise in the visual arts are no better than other participants at doing so.

2 RELATED WORK

There are three broad categories of terrain synthesis algorithm: procedural modelling, physically-based simulation, and example-based methods [22, 49].

Research in terrain synthesis initially focused on developing *procedural modelling* algorithms that approximate the shape of terrain, inspired by Mandelbrot's research that models natural phenomena using fractals [33–35]. Other procedural approaches followed, including the use of alternative fractal generation techniques [1, 19, 32, 36], distance-based functions [23, 47, 52], and sketching methods [2, 3, 7, 27]. These algorithms were designed to be fast and were especially popular due to the lack of computing power at the time. While later procedural methods allowed a greater degree of control, overall there is little to no consideration of the natural phenomena that shape terrain in the real world. Sketching approaches, in particular, are so unconstrained that they rely almost entirely on the user's knowledge of physical geography to synthesize any kind of realistic structure.

A second approach is *physically-based simulation*, using models derived from physical geography [11, 38, 39]. Some methods in this category synthesize terrain by running simulations at interactive rates [4, 5, 41, 53], and others through the use of high level evaluations [12, 40]. While these methods produced more realistic terrain than the procedural methods, most approaches rely on models of hydraulic and thermal (diffusive) erosion, ignoring other influential forces that shape terrain including glaciers, earthquakes, tectonic uplift, weather patterns, and animal and human interference. Control is also an issue, as the user can often only change the parameters of the models but cannot specify the shape of desired terrain. As a result many methods are limited in the types of terrain can be synthesized and the output can be hard to control.



(a) Tasse (b) Gain (c) Scott (d) Real (target) (e) Real (source)

Fig. 1. (a)–(c) example output of three example-based terrain synthesis methods; (d) the real terrain used as reference in this example (Set 2 in our main experiment); (e) the source terrain from which samples are drawn to generate (a)–(c).

Our work assesses the realism of *example-based methods*, which synthesize terrain from existing data. These were developed as increasing amounts of real-world elevation data became available [56]. Most of the example-based methods use techniques from texture synthesis algorithms [9, 14, 15, 17, 21, 54, 59] while others use artificial intelligence techniques [25, 48].

Figure 1(a)–(c) shows the output of three example-based methods. Each algorithm requires *source* data from which it can draw samples of realistic terrain (in this case, Figure 1(e)), and a *target* that it is trying to match. For our experiments, we created target inputs, specific to each method, that directed the method to attempt to match the overall structure of a real terrain (in this case, Figure 1(e)), so that we could ask participants to compare the generated terrains not just against each other but also against a similar real-world exemplar.

Example-based methods allow the user to have a greater degree of control over the synthesis output. However, while the local topology is realistic due to the real-world data provided, the global structure is less realistic than simulation methods as there is no consideration of natural processes. For instance, numerous endorheic basins (an area with no outflow to external bodies of water) commonly occur in a single synthesis as a result of the lack of this consideration, whereas it is rare to have even a single endorheic basin in a section of real-world terrain.

Galin et al. [22, Sec 5.1], in their survey of the field, cite three previous example-based terrain synthesis methods that build on texture synthesis: those of Zhou et al. [59] and Tasse et al. [54], based primarily on *patch-based* texture synthesis method of Wu and Yu [58], and the method by Gain et al. [21], based primarily on *pixel-based* methods of Lefebvre and Hoppe [31] and Han et al. [26].

Our work compares the output of four example-based methods: we use the methods of Tasse et al. [54] and Gain et al. [21] as the best examples of patch-based and pixel-based terrain synthesis. Our third method is a novel terrain synthesis method developed by Scott [49, 50], which builds on the *texture optimization* approach of Kwatra et al. [30], with the addition of physically-based simulation for depression breaching [51] to remove unrealistic pits and endoheric basins. Given that Scott's method uses a physically-based addition, our fourth method is our own addition of depression breaching to the method of Gain et al. [21], which allows us to ascertain whether the physically-based depression breaching makes a significant difference to the realism of Gain's method.

2.1 Realism and Believability

Realism is important in terrain synthesis, but the term 'realism' is used ambiguously. 'Realism' can be used to describe whether a terrain adheres to forms that could be generated by a real-world process but the term is often

used as well, or instead, to describe the visual aesthetics of the terrain. Conflating the two meanings causes confusion. We attempt to avoid this confusion by distinguishing between 'realism', an objective measure, and 'believability', a subjective measure.

Although there have been some prior experimental studies of subjective believability [21, 54], the overwhelming majority of terrain synthesis methods have never undergone any *experimental* evaluation of either realism or believability: instead they rely on the subjective, informal, and potentially biased evaluation by a few individuals. Where terrain methods have evaluated the 'realism' of terrain (through a subjective evaluation or otherwise), they have always been measuring the believability of the terrain, not the realism.

2.2 Previous studies in evaluating realism

Computer graphics has always attempted to emulate the real world. The Cornell Box [24] was one of the earliest examples of an attempt to compare graphics against reality. Rademacher et al. [44] considered which features, in very simple images, contribute to realism. They found statistical significance in shadow smoothness and in the roughness of surfaces, but not in the number of lights or objects. Of relevance to our study is that they compared real images against one another in assessing realism, as we do with real terrains in our first experiment, showing that equally realistic images (being of the real world) have different levels of believability. Elhelw et al. [18] comment that perception of visual realism is not well-understood and that there are many confounding factors in psychovisual experiments, including interactions between the dimensions of the visual experience, lack of evidence of what the participants are looking for, and participants' own understandings of 'realism'. Our second experiment collected data on what participants were using to judge 'realism', which we provide in our supplementary material. Kolář et al. [29] undertake a subjective evaluation of texture synthesis methods. They point to a small number of previous comparative studies on tone mapping, image retargeting, and deblurring. They use a ranking method (cf. our paired comparison method) that allows an overall judgement of which methods are statistically significantly different from one another, similar to that which we produce in our Figure 5. Beneš et al. [6] assess the realism of procedural models in architecture. They note that there is a lack of systematic treatment and understanding of what is considered 'realistic' in procedural modelling and that relatively few studies investigate realism of content in computer graphics.

The overwhelming majority of terrain synthesis algorithms have not had their realism validated through experimental study. Realism is a major part of an ideal terrain synthesis algorithm [49]. It is thus important to provide some evidence to support the claim that a terrain synthesis method produces realistic terrain.

While there is a lack of experimental evaluation of realism, there have been many studies that include experimental evaluation of quantitative data such as time-efficiency and space-efficiency, and qualitative data such as user-experience. Without a quantitative evaluation of realism, the assessment of realism (or, rather, of believability) is performed visually by the researchers. There are obvious limitations to this assessment: it is informal and unstructured, with no way to reproduce the results; researchers are motivated to develop a method that looks as realistic as possible *to them*, so are not best placed to make an impartial judgement; researchers are often not experts in the field of physical geography; and the sample size of participants is usually small.

While these issues are prevalent in the majority of studies, there are exceptions that attempt to address them. Cordonnier et al. [13] consulted an geologist who helped validate their algorithm's design. Cordonnier et al. validate the realism of their results by visually comparing the terrain profile between the method and a geological simulation; comparing various numerical aspects of the terrain, such as fold direction and wavelengths, between the method and a geological simulation; visually comparing the results of the method with various real-height-maps sourced from the USGS [56]; and collecting comments of the believability in a qualitative survey given to users after they used the method. This validation is much more credible than methods that rely solely on the visual assessment by the researchers, but it still falls short of a quantitative analysis of realism.

Reinhard et al. [46] presented some ideas on assessing parametric terrain, but did not address the question of what is considered realistic. For example, all of the fractal terrains in their Figure 4 might be considered realistic under certain conditions.

Rajasekaran et al. [45] report an experiment leading to a metric, PTRM, for the realism of terrain. Their experimental work was conducted at the same time as ours. Their work investigates *perceived realism*, a similar concept to our *believability*.. They evaluate a number of terrain generation methods, both procedural and physically-based, though all lack the sharp features that appear in the real terrains against which they compare the synthetic results, and the colouring of the terrains might have affected viewer's interpretation (a challenge with which we also grappled). Their results show that the particular synthetic terrains that they chose are statistically significantly perceived worse than real terrains. While it would be interesting to extend this method to example-based terrain synthesis, that is not how we approached our own experiments.

To our knowledge there are only two studies that have conducted a subjective evaluation of the believability of example-based terrain synthesis methods: Tasse et al. [54] and Gain et al. [21].

Tasse et al. [54] had participants compare terrain created by their system T_{sys}, against a previous related terrain deformation method [20] T_{def}, and real-world terrain T_{real}. Their hypotheses was that T_{real} is more believable than both T_{sys} and T_{def}, and that T_{sys} is more believable than T_{def}.

They found that T_{sys} was more believable than T_{def} at the 95% confidence level ($p < 0.001$), but they found no evidence that there was a difference in believability between T_{sys} and T_{real} at the 95% confidence level ($p = 0.981$). They concluded that terrain generated through their method is not dissimilar to real terrain.

This study has limitations. The terrain T_{def} and T_{sys} are dissimilar to the corresponding T_{real} (as shown in Figure 8 of Tasse et al. [54]). Our first experiment shows that this creates a 'feature bias' that confounds the results. Further, the statistical analysis was performed on an aggregation of data for the three sets instead of separately for each set. Our second experiment shows that such aggregation can confound the results.

Gain et al. [21] had participants compare real-world terrain against terrain synthesized by their method. The hypothesis was that if the synthesized terrains are as realistic as real-world terrain, then the Bernoulli distribution of the results should be similar to a sequence of coin flips.

The binomial test indicated that the proportion of synthesized terrain being more believable than real-world terrain, 0.546, was higher than the expected 0.5 ($p = .0002$, 1-sided). From this result, the coin flip hypothesis was rejected showing that synthetic terrains were actually considered more believable than the real-world terrains. Through a post-experiment questionnaire, Gain et al. [21] attributed the unusual result to subjects misidentifying prominent features in the real-world terrains, such as sharp ridges and river bends, as unrealistic. There was also anecdotal evidence that subject experts (those more familiar with physical geography) were more likely to distinguish between real-world and synthetic terrain.

This study has the same limitations as Tasse et al. [54]: the compared terrains are dissimilar in structure, creating a 'feature bias' and the analysis is performed on an aggregation of votes, confounding the results.

3 TERRAIN FEATURE BIAS EXPERIMENT

The believability of terrain may be affected by the particular features in that terrain. Our first experiment tests whether there is bias, using pair-wise comparisons of a set of real terrains. Our experimental results show a clear 'feature bias', where participants judge certain real world terrains to be more believable than other real world terrains. This has substantial implications for the design of experiments on terrain believability. This section describes that experiment and its results. It also summarises the statistical model (Bradley-Terry) that we use here and in our later experiments.

As we defined in Section 2.1, believability is not an absolute measure of realism, it is the perception of realism evaluated by an observer. For a test to evaluate the relative realism through believability, the test must be

(a) (b) (c) (d) (e) (f) (g)

Fig. 2. Terrain used in the feature bias study. (a) Mounts Sundoro and Sumbing, Central Java, Indonesia. (b) Serranía del Aguaragüe, Tarija, Bolivia. (c) West Făgăraş Mountains, Romania. (d) Sand dunes in the Southern El Djouf region, Sahara Desert, Mali. (e) Northwest of Gobernador Gregores, Santa Cruz Province, Argentina. (f) West of Mount Woodroffe (Ngarutjaranya), Australia. (g) Fractional Brownian motion created with Perlin noise. The stimuli that were presented to participants comprised these 2D top-down views plus a 3D perspective view—see the supplementary material for images.

constructed in a way that for any pair of terrains being compared, T_i and T_j, that if one is as real or more realistic than the other, $\mathfrak{R}(T_i) \geq \mathfrak{R}(T_j)$, then the same should hold true for believability, $\mathfrak{B}(T_i) \geq \mathfrak{B}(T_j)$. The validity of a test is compromised if this is not true when the test relies on assumption that believability is indicative of realism.

Feature bias is an effect where the terrain features of a height-map have an effect on the believability of the terrain independent of the realism. For example, 'typical terrain' features such as volcanic cones and dendritic river networks make the terrain more believable than terrain that is absent of those features, even if the terrains are equally realistic. That is, if $\mathfrak{R}(T_i) = \mathfrak{R}(T_j)$ and T_i contains 'typical terrain' but T_j does not, then feature bias would create an effect such that $\mathfrak{B}(T_i) > \mathfrak{B}(T_j)$. This feature bias would therefore invalidate a test where the assumption is made that believability is indicative of realism.

3.1 Experimental Design

The study is a repeated measures (within-groups) design where the participants performed a forced choice multiple paired comparison task through an online survey (which can be found in the supplementary material). The task involved being presented with a random sequence of 21 pairs of terrain images and selecting the terrain in each pair that was the 'most realistic' (where we use the word 'realistic' in the task as being more understandable to participants than the more accurate 'believable'). The 21 pairs of images consisted of all combinations of size two, from a set of seven terrain images (shown in Figure 2): three real height-maps that contain 'typical terrain' features, three real height-maps that do not contain 'typical terrain' features, and one height-map synthesized using a fractal-based method, which was chosen to generate terrain that a professional geographer would clearly see as unrealistic but which might be acceptable to a naïve observer. The six real terrain height-maps were selected as a diverse range of terrains created by various natural processes. The decision of what constitutes 'typical terrain' is subjective. All examples are at the same resolution but the volcanoes and young mountain ranges have structures that are larger and more prominent than the dunes and eroded old mountains. We hypothesized that the height-maps that contained more 'typical terrain' features (recognizable features such as volcanic cones, or dendritic river networks) would be more believable than height-maps that contained less or no 'typical terrain' features. We also hypothesized that the synthesized height-map would be more believable than terrain that contained no 'typical terrain' features.

Terrain visualisation — There are many ways to visualise terrain. We evaluated a wide range of options (see the first author's thesis [49] for detailed discussions) and chose what we believe is the best static representation of the overall terrain. We presented two views of each terrain: an overhead 2D view and a perspective 3D view. Five examples are shown in Figure 1. A colour ramp by MOSSMAN [37] was used to create the hypsometric map, with discrete colours at equally spaced intervals of 100m, from 100m to 4800m. A shaded relief map was generated with QGIS (v3.6.0-Noosa) [42] using an azimuth of 315° and an altitude of 45°, and was multiplied (at

	More realistic than									Significance						
T_i	c	a	f	b	e	g	d	Total	π_i	c	a	f	b	e	g	d
c		21	29	30	34	36	38	188	1.0000			***	***	***	***	***
a	17		29	27	31	34	38	176	0.5007			***	***	***	***	***
f	9	9		18	31	34	37	138	0.3186	***	***			***	***	***
b	8	11	20		24	31	37	131	0.2746	***	***			**	***	***
e	4	7	7	14		29	37	98	0.1342	***	***	***	**		***	***
g	2	4	4	7	9		26	52	0.0421	***	***	***	***	***		***
d	0	0	1	1	1	12		15	0.0115	***	***	***	***	***	***	

Table 1. Results of the feature-bias study and analysis using the Bradley-Terry model with Holm–Bonferroni correction. Significance values are interpreted as the confidence that the terrains are different at a given level: '.' $p < 0.1$, '*' $p < 0.05$, '**' $p < 0.01$, '***' $p < 0.001$.

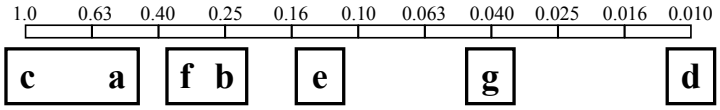


Fig. 3. A visualisation of the results of the Bradley-Terry analysis. For each set, a letter is placed at the location of the Bradley-Terry parameter. Those judged more believable are to the left of those judged less believable. Boxes are put round those letters that are *not* statistically significantly different at the 99% level. All letters within a box can be considered to be equivalent while all letters that do not share a box are statistically significantly different from one another. We use a logarithmic scale because the important feature is the relative rather than absolute magnitudes of the parameter values.

66.7% opacity) with the hypsometric map. The 2D image, was created by overlaying thin blue contours (0.5px width) generated at every 100m on the terrain texture. The 3D image was rendered as an unlit scale model with skirt that extended to sea-level and textured with the terrain texture. The camera, for the 3D view, was positioned at an altitude of 50°, a distance from the model center of 30km, and with a field of view of 60°. All terrain images can be found in the supplementary material.

Participants — We recruited 36 participants aged 19–62 (*M* 37, *SD* 10.9). Of the 36 participants, 26 were male, 11 were female, and 1 was genderqueer. All stated that they were free from any serious visual impairments. participants took part in the study willingly through an online survey and were not given any compensation for their participation.

3.2 Results

The *Total* column in Table 1 shows the number of times a terrain was evaluated to be more believable than the other terrains. This gives an ordering of believability for the terrains, where the more believable a terrain is, the larger its total of votes. However, this column cannot be taken as the definitive ordering because we need also to consider whether the differences are statistically significant. On the advice of our University's Statistical Consultant, we chose to use the Bradley-Terry model [8] for our statistical analysis.

The Bradley-Terry model is used for deriving a latent preference scale from paired comparison data, as shown in the extensive bibliography by Davidson and Farquhar [16]. The Bradley-Terry model assumes that there are

values $\pi_i \geq 0$ for each terrain T_i, such that for any pair of terrains, T_i and T_j, the probability that terrain T_i is voted more believable than terrain T_j is modelled as

$$P(\mathfrak{B}(T_i) > \mathfrak{B}(T_j)) = \frac{\pi_i}{\pi_i + \pi_j}. \tag{1}$$

The Bradley-Terry model finds the π_i values that best fit the data collected though the pair-wise comparisons and gives p-values for differences between every pair. The scale factor for the series is unimportant, so for convenience we scale the values so the largest $\pi_i = 1$. Additionally, we use Holm–Bonferroni correction [28] to control the family-wise error rate by correcting the p-values. This accounts for Type I errors when doing multiple comparisons with the same data.

The results and analysis of the pair-wise comparisons is summarised in Figure 3. This shows that there is a significant difference between five sets of terrains at the 99% confidence level.

3.3 Discussion

These results provide evidence of a feature bias that influences a participant's judgement of believability when evaluating terrain. Despite every terrain (excluding g) being equally realistic, due to them all being existing real-world terrain, there is a clear ordering of believability where terrains with more 'typical terrain' features were evaluated to be more believable than terrains without those features.

This result has substantial implications for the evaluation of synthesized terrain. Let there be two height-maps of real terrain, R_a and R_b, such that the believability of R_a is greater than R_b, ($\mathfrak{B}(R_a) > \mathfrak{B}(R_b)$). Suppose there is a data-based method that synthesizes terrain as realistic as the example terrain provided to it and it is used to synthesize two height-maps, S_a and S_b, from the example data R_a and R_b respectively. It then follows that $\mathfrak{B}(S_a) = \mathfrak{B}(R_a)$ and $\mathfrak{B}(S_b) = \mathfrak{B}(R_b)$. However if the realism of the synthesis method were to be evaluated by comparing dissimilar terrain, the conclusion may be that the synthesis method is more believable than real terrain ($\mathfrak{B}(S_a) > \mathfrak{B}(R_b)$) or less believable ($\mathfrak{B}(S_b) < \mathfrak{B}(R_a)$) depending on the comparison made in the evaluation of believability. This contradicts the actual believability of the synthesis method. Therefore, *when evaluating the relative realism of terrain, only like terrains should be compared.* In our second experiment, we control the artistic direction of the synthesis methods to avoid feature bias.

4 EXPERIMENTAL EVALUATION OF SUBJECTIVE BELIEVABILITY

This study, the main contribution of that paper, seeks to examine the believability of several data-based terrain synthesis methods by asking participants, including a substantial proportion (124 of 245 participants, 51%) who have expertise in geomorphology, to evaluate the realism of the terrain methods and real terrain using pair-wise comparisons. This study uses four methods of data-based terrain synthesis:

- Tasse (algorithm and implementation by Tasse et al. [54])
- Gain (algorithm by Gain et al. [21], implementation by us)
- Scott (algorithm by Scott [49, 50], combines texture optimization with depression-breaching)
- G-Pit (a variation of our Gain implementation to include depression-breaching [49, 50])

These four methods were chosen because Tasse and Gain are the key previous methods in example-based terrain synthesis, based on patch-based and pixel-based texture synthesis respectively. Our aim was to compare them against the new method introduced by Scott, which is based on texture optimization. However, Scott has the added feature that it includes pit removal. We therefore added a pit-removal variant of Gain. This means that we get fair comparison between Gain and Tasse (which do not have pit removal) and fair comparison between G-Pit and Scott (which do have pit removal). Furthermore, it allows us to asses whether pit-removal makes any substantial difference to the Gain algorithm by comparing Gain against G-pit.

We made an informal pilot evaluation of the methods and, from this initial evaluation, we hypothesized that the methods G-Pit and Scott would be considered more believable than the methods Gain and Tasse, and just as believable as real terrain. We also hypothesized that participants in the experiment with a greater expertise of physical geography, cartography, image analysis, or visual arts, would be able to more accurately identify real terrain from synthesized terrain than non-experts.

4.1 Design

The study used a repeated measures (within-groups) design around the pairwise comparisons of seven sets of terrain images. The decision to use seven sets was informed by the time taken for participants to evaluate terrains in the pilot study and was driven by a need to balance the number of different types of terrain against the length of the test (see Section 6.2 for a discussion of survey length). Each set used a different set of constraints and source terrain as input for each of the terrain synthesis methods (with the exception of Sets 6 and 7 which used different constraints but the same source terrain). Sets 1–5 contained five images per set (10 pairwise comparisons within each set): four created by the four methods and one of real terrain. Sets 6–7 contained four images per set (6 pairwise comparisons within each set): all created from synthesized terrains. The independent variable was a pair of terrain images to be evaluated (with 62 pairs). The dependent variable was the response of which of the terrains in the pair of terrain images was more 'realistic'. We used the word 'realistic' in the experiment, as being more easily understandable by the participants than the more accurate 'believable' (see discussion in Section 2.1). Artistic direction of the algorithms was controlled by using similar constraints for each synthesis method within each set.

We chose to use a method where, for a given set of images, every possible pair of images was compared against each other, with the participant required to select which they thought was 'most realistic' (i.e., which they found most believable). We chose to use pair-wise comparison rather than the user ranking multiple methods at once (as used by Kolář et al. [29]), as the former provides better consistency and better use of display space. We chose to include the real world example as one of the test cases (in Sets 1–5) to ascertain whether any synthetic terrain performed better than real world (as occurred in the experiment run by Gain et al. [21]); this is in contrast to the approach taken by Vanhoey et al. [57] in which they compared all stimuli against a reference.

4.2 Apparatus and Materials

The five example terrains were selected to be from sources that differ in geological structure (Table 2). The source terrain and target terrain for a given example were selected to be similar in elevation range, structure arrangement, and terrain features (see examples in Figures 1(d)–(e) and 4i). For Sets 1–5, the source, target, and synthesized height-map were all 1000×1000 pixels with a sample spacing of 1 arc-second (approximately 30 meters). Source terrain for Sets 6–7 was a resolution of 2000×2000 pixels with a sample spacing of 1 arc-second, but the synthesized height-map remained at a resolution of 1000×1000 pixels. Figure 4 shows examples of the generated terrains (the full collection of terrains is in the supplementary material).

The synthesis methods all require artistic direction. We used roughly equivalent constraints for each method. Sets 1–5 used constraints to reproduce the target real terrain in the set, Set 6 used constraints to reproduce an artificial structure similar to the one used by Zhou et al. [59] and Set 7 used no constraints. For the constraints of Tasse, we use the target terrain as input directly, as the algorithm automatically extracts the ridgelines and valley-lines it requires for the target feature-map through the use of the profile recognition and polygon breaking algorithm (PPA) [10]. Providing the target terrain directly also provides Tasse with the correct elevations and noise variance required as part of its feature-matching process. We ran the algorithm numerous times with different patch sizes (feature and non-feature) and PPA feature type (ridge, valley, or both) before selecting the most appropriate result to use in the study [49]. The other parameters were left to their default values, set by the

i	Region	Source	Target	Description of source
1	East-central Yemen	$16.167°N$ $48.139°E$	$16.167°N$ $48.417°E$	incised plateau with dry riverbeds and a high drainage density
2	Rocky Mtns, CO, USA	$39.139°N$ $106.583°W$	$39.417°N$ $106.583°W$	alpine post-glacial landscape with glacial cirque valleys and rivers that are only just starting to incise
3	Făgăraş Mtns, Romania	$45.695°N$ $24.583°E$	$45.695°N$ $24.861°E$	mix of erosional and depositional landscapes, hills transition to coalescing alluvial fans along a mountain front
4	Central Java, Indonesia	$7.389°S$ $110.083°E$	$7.500°S$ $110.445°E$	young neighbouring stratovolcanoes with a radial drainage pattern
5	Ngarutjaranya, Australia	$26.278°S$ $131.500°E$	$26.278°S$ $131.778°E$	degraded ancient landscape of inselbergs, featuring relict river valleys and recent aeolian erosion
6,7	Marlborough, New Zealand	$41.722°S$ $173.722°E$	*n/a*	tectonic modified threshold landscape, with fault bounded mountain ranges and strong parallel structural control

Table 2. Latitude-longitude coordinates, in decimal degrees, for the centre of the real-data used for the source and target terrain for each terrain Set i. Sets 6 and 7 share the same source terrain and have no real world terrain target.

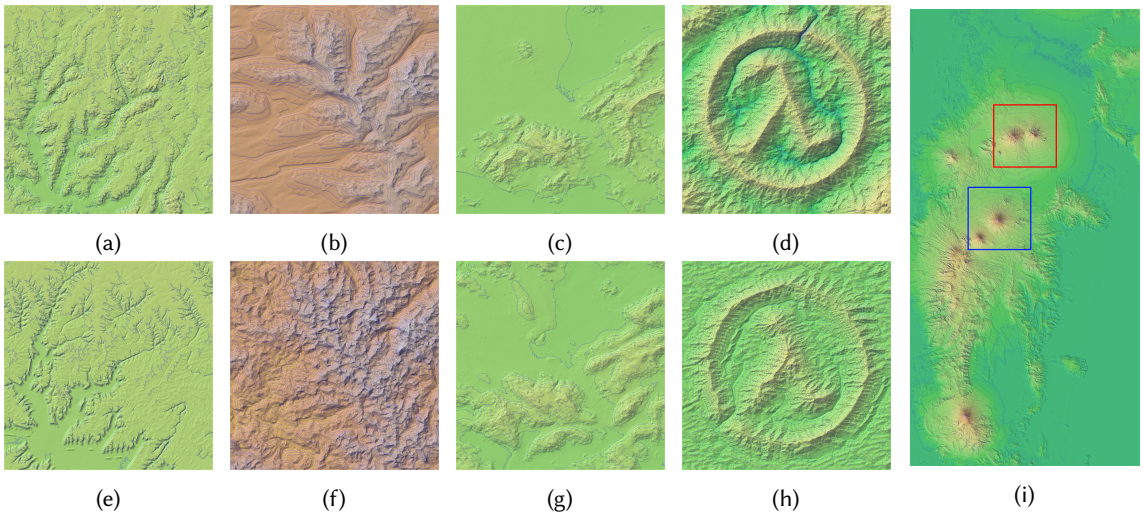


Fig. 4. The most believable (top) and least believable (bottom) examples for four of the seven sets. Most believable: (a) 1 Gain, (b) 2 Scott, (c) 5 Scott, (d) 6 G-Pit. Least believable: (e) 1 Scott, (f) 2 Tasse, (g) 5 G-Pit, (h) 6 Scott. For the full set of stimuli, see the supplementary material. (i) An example of selecting a source terrain and target terrain with similar elevation range, structure arrangement, and terrain features. The source (blue/lower) and target (red/upper) height-maps for Set 4 both contain two neighbouring stratovolcanoes with a similar layout and pattern of erosion.

original authors of the code. For the constraints of Gain, we created a program to aid in the specification of the constraints. We were able to trace the ridgelines and prominent features of the target terrain, and the program created curve constraints using the traced path, extracting the correct elevation and gradient values. We then manually corrected the area-of-effect for the constraints and other small aspects of the constraints, and the final result was a set of curve constraints that matched the target terrain. The same constraints that were created for Gain were also used for G-Pit. For the constraints of Scott, we down-sampled the target terrain four times to a

resolution of 62×62 pixels, then up-sampled it to 1000×1000 pixels, to use as the input for soft-constraints. This emulated the rough input an artist would use to guide the algorithm to produce the correct shape of terrain, without specifying specific details.

We use the same visualization method as in the previous experiment, simultaneously presenting a top-down and a perspective view of the terrain. Five examples are shown in Figure 1, with all stimuli shown in the supplementary material.

Qualtrics [43] was used to create an online survey which consisted of four sections:

(1) Participant information sheet: required by our University's Ethics Committee, and obtaining informed consent.
(2) Terrain information: how the study will be conducted, definition of realism, and an example and explanation of the terrain images that will be presented.
(3) Pair-wise comparisons: 62 pairs of stimuli broken up into eight separate pages of six to eight pairs per page. The order of and the left-right position of images in the 62 pairs were randomized. Each question presented the pair of terrain images above a horizontal radio button with three options: 'left is more realistic', 'both are equally realistic', and 'right is more realistic'. Participants could spend as long as they liked viewing the images.
(4) Participant questionnaire: questions asking for the participant's name, age, any visual impairments, and expertise in physical geography, cartography, image analysis, and visual arts. Two optional free-text questions: their strategy to determine the most realistic terrain and which image (2D overview or 3D perspective) they found most useful.

Participants required a computer with access to the internet to complete the survey, and were encouraged to complete it on a device with a large screen as opposed to a small mobile device. A copy of the survey can be found in the supplementary material. Participants were allowed as much time as they needed to complete the survey because the intention was that they have the freedom to use whatever method they like to determine which terrain appeared most 'realistic', rather than asking for a snap judgement. After completing the experiment, we collected participants' own explanations of the criteria that they had used in making their judgements.

4.3 Procedure

The survey was distributed to students and research groups at our university, local visual effects companies, and an international mailing list of professional geographers. Participants were asked to contribute to the research by completing the survey and no compensation was given. None of the researchers associated with the experiment, and no-one who had viewed the stimuli beforehand, participated in the survey. Data was collected over three weeks. We screened for visual conditions that would greatly affect the participant's ability to participate.

4.4 Participants

There were 245 valid responses. Of the completed responses there were 149 males, 18–69 years of age (M 34, SD 11.3), and 96 females, 18–68 years of age (M 33, SD 12). Six participants indicated that they had some form of red-green colourblindness but were included in the final results as it was not considered a serious enough visual impairment to be excluded.

5 RESULTS

Figure 5 summarises the analysis of comparing the seven sets individually as well as the combination of Sets 1–5 and Sets 1–7 using the Bradley-Terry method (see Section 3.2 for a summary of the Bradley-Terry statistical method). The supplementary material contains the detailed tables to support this.

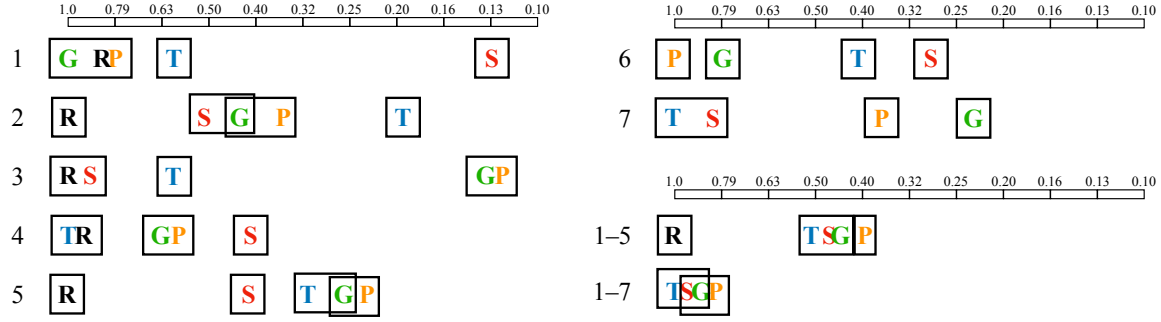


Fig. 5. A visualisation of the results of the Bradley-Terry analysis. For each set, a letter is placed at the location of the Bradley-Terry parameter, π_i. Boxes are put round those letters that are *not* statistically significantly different at the 99% level. All letters within a box can be considered to be equivalent while all letters that do not share a box are statistically significantly different from one another. R=real, T=Tasse, G=Gain, S=Scott, P=G-Pit.

From this analysis, for the five Sets (1–5) that included a real terrain, reality is always the most believable or it is statistically indistinguishable from the most believable method. This is a positive result that shows that no method has produced terrain that is more believable than reality. Each of the methods is statistically indistinguishable (at the 99% level) from reality for at least one terrain set: Tasse in Set 4, Gain and G-Pit in Set 1, and Scott in Set 3.

If we consider Gain, Scott and Tasse across the seven sets, each method performs better than each other three times, performs worse three times, and is equal once, indicating that none of the methods is consistently more believable than any other. Gain and G-Pit are almost identical: that are statistically indistinguishable at the 99% confidence level in all of Sets 1–5. Gain and G-Pit are statistically significantly different at the 99% level in the two artificial terrain sets, 6 and 7, where G-Pit is more believable. This is likely because, in Sets 6 and 7, Gain generates substantial number of pits that are removed effectively by G-Pit, whereas in Sets 1–5, the artistic direction ensures that Gain does not produce any significant pits, thereby obviating the need for the pit-removal addition. There is no one method that has generated the most believable, or least believable, terrain in all cases, rejecting our hypothesis that G-Pit and Scott would be more believable than the Tasse and Gain in all cases.

From the analysis on the aggregation of Sets 1–5 and 1–7 (where the votes are combined before the analysis), the methods are less distinguishable from one another in terms of believability. For Sets 1–5, Real is the most believable, followed by Tasse, Scott and Gain which are equally believable (at the 99% confidence level), and G-Pit as the least believable. For Sets 1–7 all four methods are almost indistinguishable from one another statistically, with the sole exception being that Tasse is more believable than G-Pit (at the 99% confidence level). This shows that the aggregation of data across the comparative terrain confounds results from the individual terrain sets. That is, there exists a clear, significant difference between methods on different types of terrain but, in aggregate, all methods are equally good.

In addition to the above analysis, we compared the accuracy of each level of expertise in the categories of physical geography G_i, cartography C_i, image analysis I_i and visual arts A_i (distributions shown in Figure 6 and a description each expertise level in Table 3). Accuracy was calculated as the fractional value of the number of times a participant indicated a real-world terrain was 'more realistic' (i.e., more believable) than a synthesized terrain, divided by the number of pairs that contained a real terrain (M 0.6963, SD 0.1635). A single factor ANOVA was computed separately for each category of expertise. For physical geographic expertise ($F(3, 241) = 13.55, p < 0.001$), cartography expertise ($F(3, 241) = 3.83, p = 0.0104$), and image analysis expertise ($F(2, 242) = 3.80, p = 0.237$), ANOVA shows that there are statistically significant differences at the $p < .05$ level.

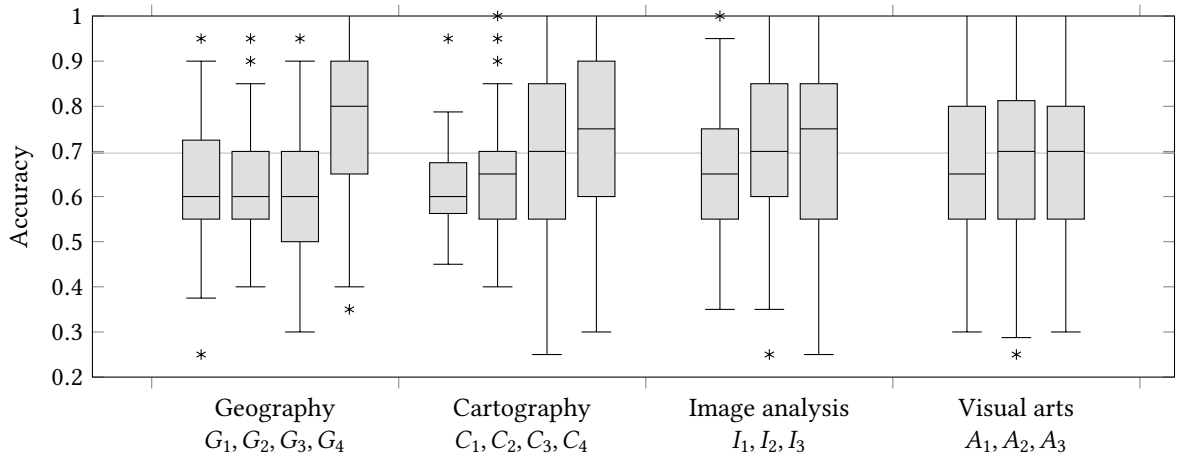


Fig. 6. A box plot of the distribution of accuracy for the independent variables of expertise in physical geography G_i, cartography C_i, image analysis I_i, and visual arts A_i. The mean accuracy for the tested population is marked with a grey line.

Physical geography (knowledge of underlying processes)	
G_1	At most, a high-school education.
G_2	Currently in a bachelor's degree or equivalent.
G_3	Completed a bachelor's degree or equivalent.
G_4	Completed a postgraduate degree or working in the field.
Cartography (ability to interpret maps)	
C_1	Cannot read a map.
C_2	Some previous experience reading a map.
C_3	Can read a map well.
C_4	An expert at reading maps.
Image analysis (experience identifying artifacts in images)	
I_1	No experience.
I_2	Some previous experience.
I_3	Regularly has to identify artifacts in images.
Visual arts (experience participating in the visual arts)	
A_1	No experience.
A_2	Some previous experience.
A_3	Regularly participates in the visual arts.

Table 3. A description of the levels for each category of expertise.

ANOVA for visual arts expertise ($F(2, 242) = 0.22, p = 0.8025$) shows that there are no statistically significant differences at the $p < .05$ level. The ANOVA tables can be found in the supplementary material.

A post-hoc analysis using Tukey's HSD test [55] revealed the following statistically significant differences at a 95% confidence level. For physical geography expertise, G_4 is statistically significantly different from any of the other three levels, but there is no statistically significant difference between any pair of the other three levels. For cartography expertise, C_4 is statistically significantly different from C_2, but that there is no statistically significant difference between any other pair of levels. The reason that C_4 and C_1 cannot be distinguished is likely owing to the small number of participants in C_1 ($N_{C1} = 10$). For image analysis expertise, I_3 is statistically significantly different from I_1, but there is no statistically significant difference between either other pair of levels. For visual arts expertise, the ANOVA showed no significant differences between any pair of levels, so it is unnecessary to run the Tukey's HSD test for this category (see supplementary material for more detail).

Out of the 245 participants, 180 provided feedback on the their strategy for determining the most believable terrain (which can be found in the supplementary material), and all 245 provided their preferred terrain image types: 30 preferred the 2D view, 108 preferred the 3D view, and 107 preferred having both the 2D and 3D view of the terrain, showing that either a 3D view, or a combination of 2D and 3D, was the most preferred visualization for the purpose of determining realism.

6 DISCUSSION

None of the tested methods could synthesize terrain that was consistently as believable as real terrain or consistently more believable than any other method's synthesized terrain. This mixed result provides evidence that making a comparison between terrain synthesis methods is a difficult task, even when many variables are accounted for. We consider two factors that might lead to this mixed result: the terrain type and the user's control over the synthesis.

To get a fair representation of the capabilities of the synthesis for each method we selected a range of different terrains that varied in geological features. The variability and complexity of each terrain likely contributed to the difference in results between the methods. For instance, in Set 1, the terrain contained a lot of ridgelines. Scott performed poorly, with the overall terrain looking blurred compared to the other methods, but Gain performed as well as real-world terrain. In comparison, in Set 3, the terrain contained a combination of hills and plains. Scott terrain performed as well as real-world terrain, but Gain performed poorly, with blurred hills that had little to no definition.

To control both the feature-bias and the artistic direction of the study, we created constraints for each method to synthesize the same structure presented in the real-world target terrain for each set. This was an important aspect of the study, as we established in Section 3 that feature-bias can severely affect the results of believability for a subjective evaluation. However, creating equal constraints for each method was challenging because the inputs for Tasse, Gain/G-Pit, and Scott all differ. There are two major issues we identified while creating the constraints for each method: over-constraining the synthesis and user expertise. Over-constraining the synthesis is an issue that applies to both Gain/G-Pit and Scott, where the number and strictness of the constraints reduces the algorithm's ability to be assessed on the believability of its output. Scott [49] suggests that the ideal terrain synthesis algorithm is one that *'produces terrain that is as real as possible, while meeting user specified constraints'*. To assess the believability of an algorithm's output, while accounting for feature-bias, an algorithm should have as few and as lenient constraints as possible, needed to replicate the structure of the target terrain. User expertise is also an issue, where the expertise of the user creating the terrains can impact the highest level of believability that can be assessed for each algorithm. The parameterization of Tasse is a challenge [49], where a large amount of the user's time is spent tuning parameters to produce the best result. We decided that we would only control the patch sizes and feature type, and leave the rest of the parameters as defaults.

6.1 Participant expertise

The results also indicate a significant difference in the accuracy (the success rate of identifying real terrains from synthesized terrain) between participants based on their levels of expertise in certain categories. In line with the anecdotal evidence provided by Gain et al. [21], participants who are experts in the field of physical geography had a higher accuracy than non-experts. However, this is only true for participants who have a post-graduate degree or who are currently working in the field of physical geography (G_4). Participants who are currently enrolled in (G_2), or have completed (G_3), a bachelor's degree or equivalent were indistinguishable from non-experts (G_1) in accuracy. Likewise, participants that are well versed in map-reading (C_4) had a higher rate of accuracy than participants with little experience reading maps (C_2). Furthermore, there was a difference in accuracy for image analysis expertise, where participants who regularly have to identify artifacts (I_3) had a higher accuracy than participants that have no experience identifying artifacts (I_1), but there was no difference between participants with different levels of expertise in the visual arts (A_1, A_2, and A_3).

We conclude that the most qualified participants to evaluate terrain are experts at the highest level in the fields of physical geography, cartography and image analysis. The mean accuracy of participants with high expertise in physical geography was the highest of any level in any group, indicating that the most qualified individuals overall to asses the realism of terrain are experts in the field of physical geography. As a result, future studies

should take into account this expertise of their participants to ensure that the results of their subjective evaluation of realism is accurately portrayed.

Experience in the visual arts is *not* an indicator that a participant is able to accurately distinguish between real and synthesized terrains. This indicates that methods that rely on the user's subjective assessment of realism are not suitable for an industry of artists (i.e., visual effects and game design) with little to no expertise in physical geography. Instead, terrain synthesis methods should be developed to ensure that the user's accuracy is irrelevant to synthesizing the most realistic terrains possible.

The written feedback regarding the participants' strategies was informative and insightful regarding what they were looking for when evaluating the realism of terrain. The complete set of feedback is in the supplementary material. We have identified three main factors that were used by many participants:

Consistency Terrain features that were unrelated, or appeared to be exact repeats, were considered unrealistic.

Detail Terrain features that were too smooth or too sharp were considered unnatural.

Drainage The presence of closed (endorheic) basins and the disconnection of drainage networks, were considered unrealistic.

Many of the participants' strategies (both experts and non-experts) involved looking at the drainage network and for the presence of closed basins, which supports our argument [50] that the presence of endorheic basins in synthesized terrains reduces the overall believability. This argument is also supported by the results for the terrain Sets 6 and 7, where pit-removal is needed more to remove the endorheic basins and synthesize more believable terrain. This is supported by the result that G-Pit was more believable than Gain in both sets, where the only difference between the methods is the integration of a pit-removal algorithm.

6.2 Length of survey

While this study had 245 participants (the largest number of participants in any survey of which we are aware in this field), it did highlight challenges with the length of the survey that we draw to the attention of those designing similar experiments. In particular, we wanted the survey to be accessible to a wide range of people around the world and to be completed by professionals who we know had limited time.

We used only static 2D and 3D views of the terrain, owing to the limitations of the survey tool we used and to allow canvassing of a wide range of participants from around the world. While it may be more appropriate to present a video of a fly-through of the terrain or to provide the participant with an interactive tool where they can explore the terrain themselves, these would have dramatically extended the time required to do the survey.

The length of the survey was carefully considered. A pilot study for the survey revealed that each pair of terrains took an average of 20 seconds to evaluate. With 62 pairs, we had an estimated completion time of just over 20 minutes (plus time to read the participant information sheet, the terrain information sheet, and fill out the participant information questions). Even with this carefully controlled time, we had as many unfinished survey responses as finished responses. Many participants complained about the length of the survey. This limits the scope that can be reasonably expected of such an experiment.

7 CONCLUSION

Our first experiment demonstrated that there is a feature bias in evaluating terrain that must be accommodated in experimental design. Our second experiment, which took that into account, was an evaluation of four example-based terrain synthesis methods: Tasse, Gain, G-Pit and Scott. We found that no method was consistently as believable as real terrain, and that no method consistently outperformed any other method. Those with expertise in the fields of physical geography, cartography or image analysis were more readily able to identify real terrain

more accurately from generated terrain. There was no indication that expertise in visual arts provided such an ability.

Our experimental work provides an exemplar for future studies of terrain realism. We have identified one confounding factor, feature bias, that needs to be taken into account when assessing terrain. We note that our method of presenting terrains, as static images viewed from a distance, is rather different from the use case in games, where terrain is viewed close up. This does raise questions of whether players of a game would notice the unrealistic features that these methods generate, such as unrealistic pits or overly smooth mountains. We have left open which way of presenting terrain is most appropriate: we used static images in order to allow us to access a large, expert group of participants, but future work could include an exploration of dynamic stimuli, whether video or participant-guided, and compare this to static imagery.

Our second experiment showed that all four of the tested methods can perform well and all four can perform poorly. Performance depends on the type of terrain being generated and different methods excel at different types. There is much scope for assessing and improving example-based terrain synthesis to see if it is possible to create a method that works well on a wide range of terrain types. There is also scope [45] to develop methods that can automatically assess the subjective quality of terrain. Further analysis of our participants' strategies (Section 6.1) could inform such development.

ACKNOWLEDGMENTS

Scott was supported by a Wellington Doctoral Scholarship from Victoria University of Wellington.

REFERENCES

[1] Adams, D., Egbert, P., and Brunner, S. Feature-based interactively sketched terrain. In *Proc. ACM SIGGRAPH Symp. on Interactive 3D Graphics and Games* (2012), I3D '12, ACM, pp. 208–208.

[2] Becher, M., Krone, M., Reina, G., and Ertl, T. Feature-based volumetric terrain generation. In *Proc. 21st ACM SIGGRAPH Symp. on Interactive 3D Graphics and Games* (2017), I3D '17, ACM, pp. 10:1–10:9.

[3] Belhadj, F., and Audibert, P. Modeling landscapes with ridges and rivers: Bottom up approach. In *Proc. 3rd Int. Conf. on Comp. Graphics and Interactive Techniques in Australasia and South East Asia* (2005), GRAPHITE '05, ACM, pp. 447–450.

[4] Beneš, B. Physically-based hydraulic erosion. In *Proc. 22nd Spring Conference on Computer Graphics* (2006), SCCG '06, ACM, pp. 17–22.

[5] Beneš, B., Těšìnskỳ, V., Hornyš, J., and Bhatia, S. K. Hydraulic erosion. *Computer Animation and Virtual Worlds 17*, 2 (2006), 99–108.

[6] Beneš, J., Kelly, T., Děchtěrenko, F., Křivánek, J., and Müller, P. On realism of architectural procedural models. *Comp. Graphics Forum 36*, 2 (2017), 225–234.

[7] Bernhardt, A., Maximo, A., Velho, L., Hnaidi, H., and Cani, M.-P. Real-time terrain modeling using CPU-GPU coupled computation. In *Proc. 24th SIBGRAPI Conference on Graphics, Patterns and Images* (2011), SIBGRAPI '11, IEEE Computer Society, pp. 64–71.

[8] Bradley, R. A., and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika 39*, 3/4 (1952), 324–345.

[9] Brosz, J., Samavati, F. F., and Sousa, M. C. Terrain synthesis by-example. In *Advances in Computer Graphics and Computer Vision: International Conferences VISAPP and GRAPP 2006*, J. Braz, A. Ranchordas, H. Araújo, and J. Jorge, Eds. Springer, 2007, pp. 58–77.

[10] Chang, Y.-C., Song, G.-S., and Hsu, S.-K. Automatic extraction of ridge and valley axes using the profile recognition and polygon-breaking algorithm. *Computers & Geosciences 24*, 1 (1998), 83–93.

[11] Chiba, N., Muraoka, K., and Fujita, K. An erosion model based on velocity fields for the visual simulation of mountain scenery. *The Journal of Visualization and Computer Animation 9*, 4 (1998), 185–194.

[12] Cordonnier, G., Braun, J., Cani, M.-P., Beneš, B., Galin, E., Peytavie, A., and Guérin, E. Large scale terrain generation from tectonic uplift and fluvial erosion. In *Comp. Graphics Forum* (2016), vol. 35, pp. 165–175.

[13] Cordonnier, G., Cani, M.-P., Beneš, B., Braun, J., and Galin, E. Sculpting mountains: Interactive terrain modeling based on subsurface geology. *IEEE Trans. Visualization and Comp. Graphics 24*, 5 (May 2018), 1756–1769.

[14] Cruz, L., and Velho, L. *High-Level Techniques for Landscape Creation.* PhD thesis, Instituto Nacional de Matemática Pura e Aplicada, Rio de Janeiro, Brazil, Mar 2015.

[15] Cruz, L., Velho, L., Galin, E., Peytavie, A., and Guérin, E. Patch-based terrain synthesis. In *Proc. 10th Int. Conf. on Comp. Graphics Theory and Applications* (Berlin, France, 2015), GRAPP.

[16] Davidson, R. R., Farquhar, P. H., et al. A bibliography on the method of paired comparisons. *Biometrics 32*, 2 (1976), 241–252.

[17] dos Passos, V. A., and Igarashi, T. LandSketch: A first person point-of-view example-based terrain modeling approach. In *Proc. Int. Symp. on Sketch-Based Interfaces and Modeling* (2013), SBIM '13, ACM, pp. 61–68.
[18] Elhelw, M., Nicolaou, M., Chung, A., Yang, G.-Z., and Atkins, M. S. A gaze-based study for investigating the perception of visual realism in simulated scenes. *ACM Trans. Applied Perception 5*, 1 (2008), 1–20.
[19] Fournier, A., Fussell, D., and Carpenter, L. Computer rendering of stochastic models. *Commun. ACM 25*, 6 (June 1982), 371–384.
[20] Gain, J., Marais, P., and Strasser, W. Terrain sketching. In *Proc. 2009 Symp. on Interactive 3D Graphics and Games* (2009), I3D '09, ACM, pp. 31–38.
[21] Gain, J., Merry, B., and Marais, P. Parallel, realistic and controllable terrain synthesis. *Comp. Graphics Forum 34*, 2 (2015), 105–116.
[22] Galin, E., Guérin, E., Peytavie, A., Cordonnier, G., Cani, M.-P., Benes, B., and Gain, J. A review of digital terrain modeling. *Comp. Graphics Forum 38*, 2 (2019), 553–577.
[23] Golubev, K., Zagarskikh, A., and Karsakov, A. Dijkstra-based terrain generation using advanced weight functions. *Procedia Computer Science 101* (2016), 152 – 160. 5th Int. Young Scientist Conf. on Computational Science, YSC 2016, 26-28 October 2016, Krakow, Poland.
[24] Goral, C. M., Torrance, K. E., Greenberg, D. P., and Battaile, B. Modeling the interaction of light between diffuse surfaces. *Computer Graphics (Proc. SIGGRAPH) 18*, 3 (1984), 213–222.
[25] Guérin, E., Digne, J., Galin, E., Peytavie, A., Wolf, C., Beneš, B., and Martinez, B. Interactive example-based terrain authoring with conditional generative adversarial networks. *ACM Trans. Graph. 36*, 6 (Nov. 2017), 228:1–228:13.
[26] Han, C., Risser, E., Ramamoorthi, R., and Grinspun, E. Multiscale texture synthesis. In *SIGGRAPH 2008* (2008), ACM, pp. 51:1–51:8.
[27] Hnaidi, H., Guérin, E., Akkouche, S., Peytavie, A., and Galin, E. Feature based terrain generation using diffusion equation. In *Comp. Graphics Forum* (2010), vol. 29, pp. 2179–2186.
[28] Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.
[29] Kolář, M., Debattista, K., and Chalmers, A. A subjective evaluation of texture synthesis methods. *Comp. Graphics Forum 36*, 2 (2017), 189–198.
[30] Kwatra, V., Essa, I., Bobick, A., and Kwatra, N. Texture optimization for example-based synthesis. In *SIGGRAPH 2005* (2005), ACM, pp. 795–802.
[31] Lefebvre, S., and Hoppe, H. Appearance-space texture synthesis. In *SIGGRAPH 2006* (2006), ACM, pp. 541–548.
[32] Lewis, J. P. Generalized stochastic subdivision. *ACM Trans. Graph. 6*, 3 (July 1987), 167–190.
[33] Mandelbrot, B. B. Stochastic models for the Earth's relief, the shape and the fractal dimension of the coastlines, and the number-area rule for islands. *Proc. National Academy of Sciences 72*, 10 (1975), 3825–3828.
[34] Mandelbrot, B. B. *The Fractal Geometry of Nature*. W. H. Freeman and Company, 1977.
[35] Mandelbrot, B. B. Fractal landscapes without creases and with rivers. In *The Science of Fractal Images*, H.-O. Peitgen and D. Saupe, Eds. Springer-Verlag, 1988, pp. 243–260.
[36] Miller, G. S. P. The definition and rendering of terrain maps. In *SIGGRAPH 1986* (1986), ACM, pp. 39–48.
[37] Mossman, J. New color system enhances relief maps. https://www.esri.com/news/arcuser/0101/shademax.html Accessed: 28-06-2019.
[38] Musgrave, F. K., Kolb, C. E., and Mace, R. S. The synthesis and rendering of eroded fractal terrains. In *SIGGRAPH 1989* (1989), ACM, pp. 41–50.
[39] Nagashima, K. Computer generation of eroded valley and mountain terrains. *Visual Computer 13*, 9 (1998), 456–464.
[40] Natali, M., Lidal, E. M., Parulek, J., Viola, I., and Patel, D. Modeling terrains and subsurface geology. In *Eurographics 2013 State of the Art Reports (STARs)* (2013), Eurographics, pp. 155–173.
[41] Neidhold, B., Wacker, M., and Deussen, O. Interactive physically based fluid and erosion simulation. In *Proc. First Eurographics Conf. on Natural Phenomena* (Aire-la-Ville, Switzerland, 2005), NPH'05, Eurographics, pp. 25–33.
[42] QGIS Development Team. *QGIS Geographic Information System*. Open Source Geospatial Foundation, 2009.
[43] Qualtrics Labs, Inc. Qualtrics, 2009. https://www.qualtrics.com Accessed: 28-06-2019.
[44] Rademacher, P., Lengyel, J., Cutrell, E., and Whitted, T. Measuring the perception of visual realism in images. In *Eurographics Workshop on Rendering Techniques* (2001), Springer, pp. 235–247.
[45] Rajasekaran, S. D., Kang, H., Čadík, M., Galin, E., Guérin, E., Peytavie, A., Slavík, P., and Benes, B. PTRM: Perceived terrain realism metric. *ACM Trans. Appl. Percept.* (Jan 2022). In Press.
[46] Reinhard, E., Shirley, P., Ashikhmin, M., and Troscianko, T. Second order image statistics in computer graphics. In *Proc. First Symp. on Applied Perception in Graphics and Visualization* (2004), pp. 99–106.
[47] Rusnell, B., Mould, D., and Eramian, M. Feature-rich distance-based terrain synthesis. *Visual Computer 25*, 5-7 (Apr. 2009), 573–579.
[48] Saunders, R. L. Realistic terrain synthesis using genetic algorithms. Master's thesis, Texas A&M University, 2006.
[49] Scott, J. J. *Realism in Data-based Terrain Synthesis*. PhD thesis, Victoria University of Wellington, April 2020. http://hdl.handle.net/10063/9025.
[50] Scott, J. J., and Dodgson, N. A. Example-based terrain synthesis with pit removal. *Computers & Graphics 99* (2021), 43–53.
[51] Soille, P. Morphological carving. *Pattern Recognition Letters 25*, 5 (2004), 543–550.
[52] Stachniak, S., and Stuerzlinger, W. An algorithm for automated fractal terrain deformation. In *Proc. Comp. Graphics and Artificial*

Intelligence (2005), vol. 1, pp. 64–76.

[53] Šťava, O., Beneš, B., Brisbin, M., and Křivánek, J. Interactive terrain modeling using hydraulic erosion. In *Proc. 2008 ACM SIGGRAPH/Eurographics Symp. on Comp. Animation* (Aire-la-Ville, Switzerland, Switzerland, 2008), SCA '08, Eurographics, pp. 201–210.

[54] Tasse, F. P., Gain, J., and Marais, P. Enhanced texture-based terrain synthesis on graphics hardware. *Comp. Graphics Forum 31*, 6 (2012), 1959–1972.

[55] Tukey, J. W. Comparing individual means in the analysis of variance. *Biometrics 5*, 2 (1949), 99–114.

[56] USGS. Shuttle radar topography mission, 2006. College Park, Maryland, February 2000.

[57] Vanhoey, K., Sauvage, B., Kraemer, P., and Lavoué, G. Visual quality assessment of 3D models: on the influence of light-material interaction. *ACM Trans. Applied Perception 15*, 1 (2017), 1–18.

[58] Wu, Q., and Yu, Y. Feature matching and deformation for texture synthesis. In *SIGGRAPH 2004* (2004), ACM, pp. 364–367.

[59] Zhou, H., Sun, J., Turk, G., and Rehg, J. M. Terrain synthesis from digital elevation models. *IEEE Trans. Visualization and Comp. Graphics 13*, 4 (July 2007), 834–848.