GI_Forum 2021, Issue 2
Page: 174 - 192
Full Paper
Corresponding Author:
christoph.traun@sbg.ac.at
DOI: 10.1553/giscience2021_02_s174

Design of an Experiment to Evaluate Modes of Value Generalization in Animated Choropleth Maps

Christoph Traun, Gudrun Wallentin and Manuela Larissa Schreyer
Salzburg University, Austria

Abstract

In this research, we discuss five design parameters found in cognitive studies related to animated choropleth maps and compare existing studies accordingly. With reference to these parameters, we present the design of an experiment to assess how different forms of value generalization in choropleth map animation affect the perception of overall trends and local deviations therefrom. The mixed study design of the experiment allowed a direct comparison of within-subject and between-subject designs. Using repeated subsampling of our empirical data, the greater statistical power of the within-subject design was proven, even when correcting for differences in the number of data elements analysed.

Keywords:

cartography, empirical study design, choropleth map animation

1 Introduction

In these times of the COVID-19 pandemic, informational dashboards are mushrooming around the globe. Many of them contain temporal animations of choropleth maps, showing the development of such things as the number of infections per 100,000 people over time. Whether or not choropleth map animation is the best choice for such an endeavour, it is at least a popular way to depict time-series data aggregated to enumeration units, as it shows the dynamics of spatial processes in a direct fashion (Campbell & Egbert, 1990; Multimäki, 2016) and is technically straightforward to implement (Butler, 2016).

As animated maps pose a heavy processing workload on the human brain (Harrower, 2007a), generalization is expected to help in the pattern recognition process (Harrower, 2003; Monmonier, 1996). Thus, we conducted an experiment to investigate the effect of value generalization ('smoothing of values/colours') on the perception of choropleth map animations. In order to design this experiment, we surveyed the literature on cognitive studies on animated choropleth maps. Despite the popularity of map animation, related empirical research is relatively sparse and results are difficult to judge and compare due to the great variety of study setups used. This sparked our interest and we identified several fundamental design criteria connected to this type of empirical cartographic research, namely

- the type of data used for stimulus design,
- the number of frames per stimulus,
- frame duration,
- the number of stimuli exposures per person, and
- the (statistical) design of the study

While all the studies reviewed comment on one or other of these criteria and decisions on the respective parameters, there are several instances where important choices are barely justified or are seemingly made for the sake of convenience or pragmatism. As the interpretation of data from experiments involving humans depends greatly on the design of the test instruments being used (Olson, 2009), we consider it worth shedding light on this important aspect of cognitive cartographic research in general and as it relates to choropleth map animation in particular.

In the following section, we briefly discuss the criteria mentioned in the context of the empirical studies identified. To exemplify the rationales associated with the choice of particular parameters, we present the design of an experiment that aims to assess the impact of value generalization in time, in space, and in a combination of space and time, on a person's ability to determine global trends as well as to detect local outliers from them.

While the discussion from a cognitive perspective of the data acquired is outside the scope of this paper, a mixed within-subject and between-subject design allows us to perform a comparative power analysis of those study designs that are widely used in empirical cognitive (cartography) research –something that we had not encountered in our domain.

2 Design criteria for empirical studies on the cognition of animated choropleth maps

From among the numerous empirical studies involving map animation (see Lobben (2008) for a taxonomy of evaluation-based approaches in the map animation literature), we surveyed studies focussing on cognitive effects related to variation of properties (data structure, map design) of animated choropleth maps. In Table 1, we list all such studies that we are aware of, and compare them by design criteria that we consider worthy of discussion.

Table 1: Cognitive studies on animated choropleth maps

Author and year	Aim of the study	Type of data used	# of frames per stimulus	Frame duration in ms	# of stimulus exposures per person	(Statistical) study design
(A. L. Griffin, MacEachren, Hardisty, Steiner, & Li, 2006)	Identification of moving clusters in animated maps and small multiples	Synthetic data – regular array of 756 hexagons	6	250 350 450 550	26	Within-subject design – 24 students
(Harrower, 2007b)	Comparison of classed and unclassed choropleth map animation	Unemployment rates for >3100 US counties	34	125	2	Within-subject design – 55 students
(McCabe, 2009)	Map reading task performance between raw, temporally averaged, and temporally aggregated data	Measles infection rates for 35 different-sized districts in Niger	52 (26 in aggregated version)	400	15	Between-subject design – 96 students in 3 groups
(Fish, Goldsberry, & Battersby, 2011)	Influence of the design of scene transitions on change-detection abilities	Unemployment rates for 64 counties in Georgia	2	2000	108	Within-subject design – 78 students
(DuBois, 2013)	Effects of cluster intensity, number of simultaneous clusters, and cluster position on change cluster detection	Synthetic data – 1541 (relatively similar-sized) polygons	2	2000	39	Within-subject design – 84 students
(Moon, Kim, & Hwang, 2014)	Effects of magnitude of change and spatial distribution on gross change-detection	Synthetic data – 42 (relatively similar-sized) polygons	2	1000 2000 3000	108	Within-subject design – 18 students
(Cybulski & Medyńska-Gulij, 2018)	Effects of increasing polygon border width on the identification of extreme values.	(Fictional) unemployment rates for 14 different-sized districts in Saudi Arabia	10	3000	1	Between-subject design – 60 students in 2 groups
(Cybulski & Krassanakis, 2021)	Effects of different magnitude of change conditions on correct detection of non-changing polygons	Data of unclear origin – 3 maps of 7, 24 and 54 very different-sized polygons	2	unlimited (user controlled)	3	Between-subject design – 45 students in 3 groups

2.1 Type of data

Map stimuli for empirical tests can be made from real-world geospatial data, synthetic data, or combinations of the two, like real geometries and synthetic attributes. While real-world data have the benefit of being realistic by default, synthetic data have the following advantages:

- Good control over data and therefore stimulus properties like the degree of spatial and temporal autocorrelation,
- prevention of unwanted confounding effects due to familiarity with the geographic region or topic,
- avoidance of large area-differences of enumeration units.

Referring to the latter, McCabe (2009) self-critically notes that the high visual salience of large but sparsely populated districts in Niger probably distorts the results of his study. Being aware of the problem that changes in large counties 'visually overpower other locations', Harrower (2007b) limited his questions to several, highlighted US areas containing counties of similar size. As the generation of dynamic geospatial data with the desired properties is itself challenging (A. L. Griffin et al., 2006), there might be pragmatic reasons for only three out of eight studies using artificially generated data, and except for the study by A. L. Griffin et al. (2006), even those use stimuli comprising just two frames and disregard temporal autocorrelation.

2.2 Number of frames per stimulus

Half of the studies reviewed use stimuli limited to two animation frames. This seems valid for the assessment of human ability to detect change between two map scenes while focusing on individual enumeration units (Cybulski & Krassanakis, 2021; Fish et al., 2011), clusters of units (DuBois, 2013), or the gross change for the whole map (Moon et al., 2014). From our perspective, however, the question remains open as to whether the task of viewing and understanding a map animation (portraying the fluid development of a spatial process) is cognitively equivalent to evaluating a considerable number of concatenated before/after changes within a map scene.

2.3 Animation speed

Although studies are not fully consistent (Moon et al. (2014), for example, did not find differences in correctness of response in relation to animation speed), the viewing time per frame is considered to be positively correlated to change-detection capabilities (McCabe, 2009; Multimäki & Ahonen-Rainio, 2015). Conversely, the slower an animation gets, the more it loses its potential to create an illusion of motion due to coordinated colour changes of adjacent polygons. To give an example, the outbreak of an infectious disease might lead to the impression of an expanding figure of high infection rates around its origin. If the frame rate is too low, the apparent motion of the expanding boundary cannot be established by our visual system (A. L. Griffin et al., 2006). The studies reviewed vary a lot in animation pace, from a slideshow-like 3 seconds per frame up to a speed 24 times greater, of 8 frames per second.

2.4 Number of stimulus exposures per person

The overall number of stimuli to which each participant is exposed during an empirical test depends on several criteria:

- The number of independent variables (factors) to test, e.g. effect of animation speed and generalization type.
- The number of levels for each factor, e.g. three different animation speeds.
- The number of tasks being tested, e.g. reading out values, focussing on global trends, comparing two regions.
- The risk that some unnoticed yet special configuration in one particular map stimulus acts as a confounding factor and thus leads to 'untypical' results. Thus it is wise to take such erratic behaviour into account through the redundant use of several different base stimuli.
- The overall test design, namely whether each person has to see every condition or only a subset thereof.

The range of the numbers of stimulus trials in the studies reviewed is enormous. Cybulski and Medyńska-Gulij (2018), for example, test three different map types, including a choropleth map, using just one base stimulus for each. The two levels (unmodified/enhanced borders) of the choropleth map stimulus were presented to different respondents. Harrower (2007b) circumvents the problem of having just one large stimulus map by concentrating on several subdivisions – which could be seen as redundant stimuli themselves. Conversely, Fish et al. (2011) and Moon et al. (2014) had their respondents work through 108 stimuli – and put their patience and concentration to the test.

2.5 Study design

In behavioural and cognitive sciences, there are two basic study designs: between-subject and within-subject testing (Charness, Gneezy, & Kuhn, 2012; Keren, 2014). In a typical within-subject design, each participant is exposed to different versions of the same (base) stimulus, modified by the independent variable(s). Since only relative performance differences within each person are analysed, a within-subject design controls for heterogeneity between participants and potentially different testing environments. Thus, within-subject testing is often associated with higher statistical power and smaller sample sizes. Disadvantages are decreasing motivation due to longer tests and the danger of carry-over effects. If, for example, participants remember a stimulus they have already seen, their reaction is altered and the experiment is flawed. Although attempts are made to minimize this risk by temporally separating and concealing variants of the same stimulus-maps (e.g. through rotating or mirroring the variants), between-subject designs are safe in this respect as each person – randomly assigned to one condition – is exposed to each stimulus only once. Due to increased variance caused by personal differences and the division of participants into several groups, between-subject testing requires a larger sample size. In section 4 of this article, we compare the performance of both test designs by using empirical data from our mixed (within- and between-subject) study design on generalization of animated choropleth maps, which is outlined below.

3 An experiment on value generalization in animated maps

In this section, we address some rationales of cognitive cartographic experimental design using a real-world example. The experiment was applied in a full-scale empirical study (Traun, Schreyer, & Wallentin, 2021), aiming to examine whether and how different forms of value generalization of unclassed choropleth map animations affect the ability of users to detect general trends and local outliers thereof. The experiment was in two parts. In the first part, participants are exposed to short sequences of animated map frames, with each animation ('stimulus') consisting of a general trend and two local outliers. These local outliers are polygons with values that differ greatly from the mean value of their neighbours in space AND from the mean value of their neighbours in time (for the definition of neighbourhood, see Figure 2). Immediately after each animation, the stimulus disappears and participants have to select the correct outliers from a set of outlier candidates. In the second part, the same stimuli are shown again, but now test subjects are told to focus on the overall trend and to choose the correct trend response item after the stimulus disappears. Using differently generalized versions of the animations (non-generalized reference, temporally-, spatially- and spatiotemporally generalized versions), we investigate the effect of the generalization mode on the ability of users to correctly detect local outliers and overall trends. The following subsections describe the development of the animated maps used as test stimuli (3.1), the response items participants have to choose from (3.2), the study design and implementation (3.3), and finally the dataset acquired (3.4).

3.1 Test stimuli

Each of our stimuli comprises 14 animation frames showing unclassed attribute/colour changes of an artificial basemap. The basemap consists of 85 irregular enumeration units of roughly similar size to prevent visual dominance of large polygons while preserving the familiar 'choropleth map look' (Figure 1).

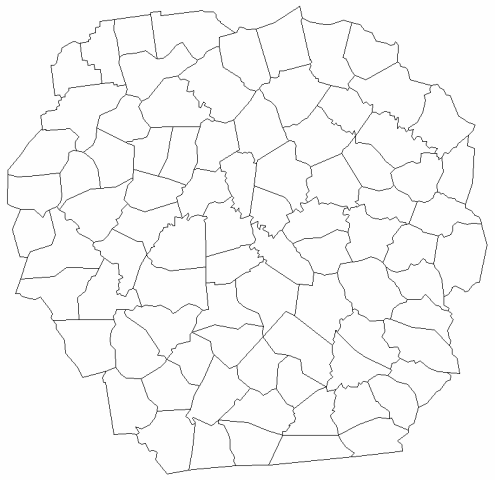

Figure 1: Basemap used for synthetic time-series data

To construct the basemap, we clipped a roughly circular subset of Counties from the US State of Kentucky and merged, split or freely changed several County geometries to approximate their areas. According to Traun and Mayrhofer (2018), choropleth map animation of time-series data seems useful only for data that exhibits a high level of autocorrelation in space and time, as this will lead to a continuous development of patterns throughout the animation. In turn, weakly autocorrelated data result in uncoordinated flicker, hampering perception, and thus calling into question the use of map animation generally. For generating such highly autocorrelated synthetic data, we developed a model using the GAMA agent-based simulation environment (https://gama-platform.github.io). The model produces moving clusters embedded in a global trend and allows for parameterizing spatial autocorrelation, temporal autocorrelation, the weight of the overall trend and clustering, and the probability of local outliers in space and time. After determining suitable parameter settings (significant clustering, high autocorrelation, rare appearance of local outliers), a large number of test stimulus candidates was simulated. From these, we chose five time-series stimuli containing exactly two local outlier polygons in space and time in order to ensure comparability between stimuli when analysing outlier detection performance. In addition, we selected only stimuli in which the two outlier polygons

- were separated by at least 0.6 seconds, in order to avoid attentional blink phenomena (Raymond, Shapiro, & Arnell, 1992), and where
- local outlier polygons were not located at the beginning or the end of the 14-frame time-series sequence.

Local outliers were determined by a high value-difference in relation to their first-order neighbourhood in space and time. To define thresholds, we used the heuristic of Traun and Mayrhofer (2018) that evaluates local differences in the context of global autocorrelation. It is based on the rationale that a certain local difference might qualify as a local outlier in a smoothly changing, highly autocorrelated dataset, but not in a less autocorrelated, 'rougher surface' (see Traun & Mayrhofer, 2018 for further detail). The explicit generation of a few (the probability of a polygon being a local outlier was set to = 0.002) but highly deviating local values within the otherwise relatively 'smooth' data from the GAMA model allowed a fairly clear separation of local outliers from 'regular' polygons.

Because of our interest in different types of generalization, we derived three differently generalized versions for each of the test stimuli by smoothing the polygon values that we generated by their

- spatial neighbours
- spatiotemporal neighbours
- temporal neighbours (Figure 2).

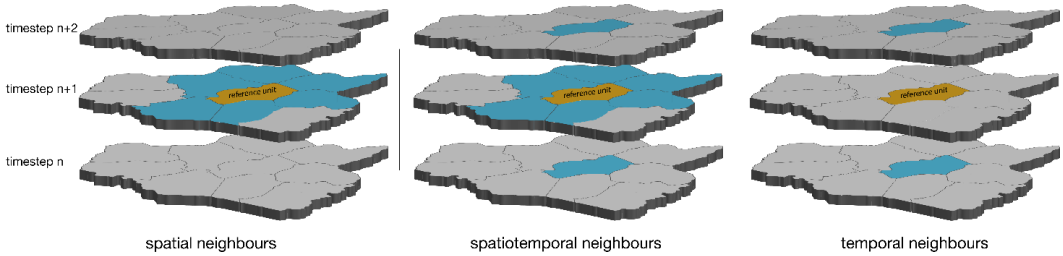

Figure 2: The three definitions of neighbourhood (blue) shown for a single reference unit (orange).

Using the methods and software provided by Traun and Mayrhofer (2018), the amount of smoothing adapts to the degree of autocorrelation in the data. Since all stimuli result from the same model parameters, spatiotemporal autocorrelation is similarly high, with Moran's *I*s (Moran, 1950) being between 0.84 and 0.93. Consequently, the strength of the value generalization is fairly similar for all stimuli. It is important to note that the values of the two local outlier polygons in each stimulus were excluded from these generalization processes.

As we wanted to avoid potential visual artefacts introduced by cartographic classification, we opted for unclassed animated choropleth maps (Harrower, 2007b) and linearly applied a sequential yellow-to-dark-brown continuous colour scheme to the data values. While the two local outliers preserve their colour throughout all variants of each stimulus, generalization shifts the colour of other polygons slightly towards the mean value of their spatial, temporal or spatiotemporal neighbours, leading to smoother overall changes within the respective dimension (Figure 3).

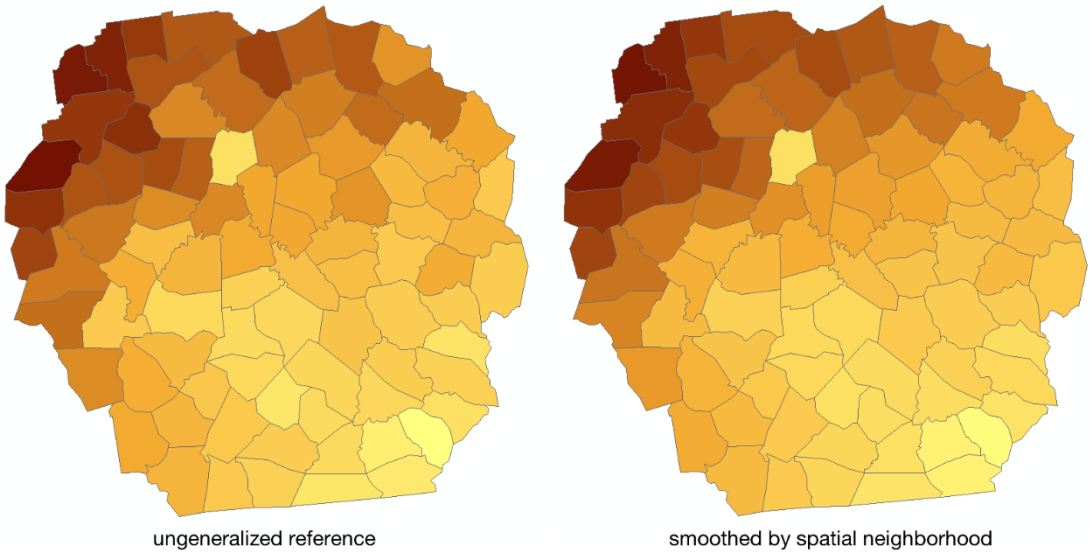

Figure 3: Subtle reduction of local contrast in the spatially generalized version (right) of a frame from Stimulus 1. The colour of a local outlier polygon (light-coloured polygon in the upper-left from centre) is not altered by generalization. Source: Traun et al. (2021)

All test stimuli, measuring 512 by 512 px, were shown at a rate of 5 frames per second and without tweening (Battersby & Goldsberry, 2010; Fish et al., 2011). The resulting 200 ms per frame are within the 'appropriate' speed range of 100 to 400 ms found by McCabe (2009), while still being both fast enough to get a passable impression of animation and slow enough to discern individual map frames (Harrower & Fabrikant, 2008). Each stimulus animation was preceded by a leader showing the empty basemap polygons superimposed by a three-second countdown text 'Start in [3, 2, 1] seconds'. While one of the five stimuli was used as a trial stimulus to familiarize participants with the experiment, four stimuli (referred to in what follows as Stimulus 1, 2, 3 and 4) were used to collect data.

3.2 Response items

Immediately after a stimulus was shown, it was replaced by a set of response items on local outliers (first part of the experiment) or global trends (second part) for participants to choose from.

Local outlier response items

Six randomly arranged basemaps, highlighting one outlier candidate each, contain the two correct outliers from the stimulus and four wrong outlier candidates (Figure 4).

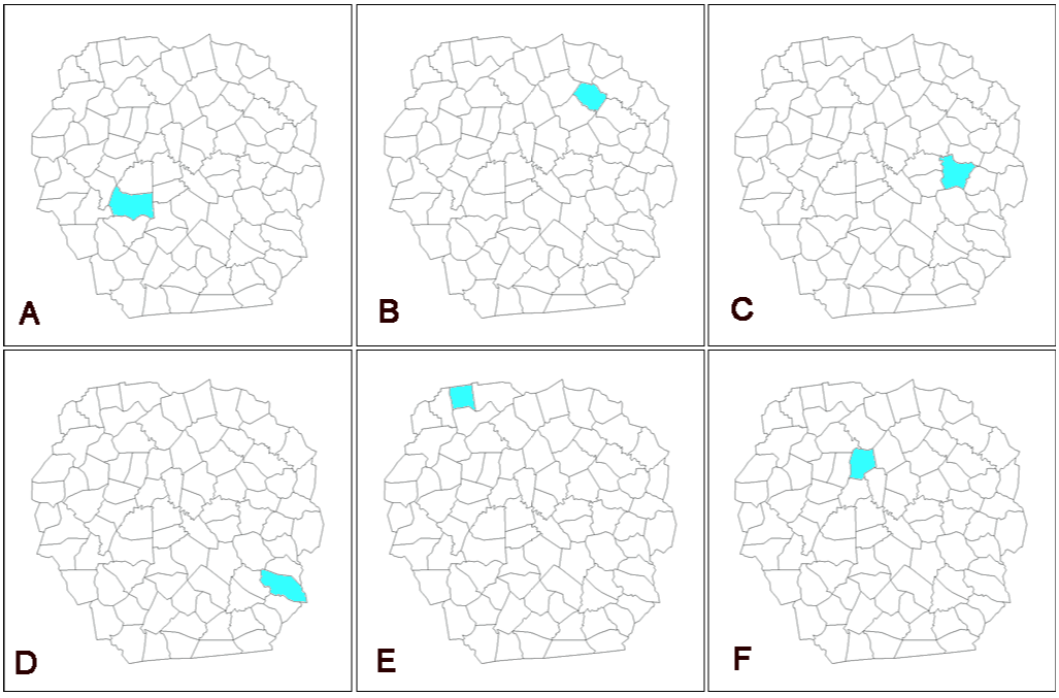

Figure 4: Local outlier response items for Stimulus 1. Compare the options A to F to the local outlier shown in Figure 3. Source: Traun et al. (2021)

For the incorrect candidates, we did not randomly choose non-outlier polygons from the basemap but founded our selection on two criteria:

- Throughout the animation, incorrect outlier candidates should have a low value difference to their spatiotemporal neighbourhood. Thus, polygons were ordered according to their maximum absolute value difference during the whole animation sequence, and incorrect outlier candidates were chosen from the lowest third of the resulting distribution (with the two real outliers being at the top of the distribution). This ensures that participants cannot choose medium local contrast polygons from the non-generalized reference animations, misinterpreting them as local outliers.
- Only polygons not adjacent to other outlier candidates (regardless of whether they were correct or incorrect) qualified as outlier candidates. Spatial dispersion of outlier candidates across the basemap was preferred to prevent participants being confused by similar location. For example, if the correct local outlier is located in the upper left of the animation, there should be just a single outlier candidate within this region.

Giving a clear separation of correct and incorrect outlier candidates both in local brightness contrast and in location, we aimed to prevent accidental misclassification. We thus increased the discriminatory power of the experiment in terms of measuring perception differences between different modes of generalization.

Global trend response items

Compared to the measurement of outlier detection capability, the development of a test for global trend detection was challenging. We soon decided against the use of verbal descriptions of the patterns seen, as this would require further abstraction, including a transformation from visual to verbal cognitive modes. In a pre-test, we tried to assess the perception of pattern-change by using blurred still-images of the first, the seventh and the last frame of the stimulus animation (true candidate), and two other animations (false-trend candidates) (Figure 5). Low success rates, at times close to a random choice, indicated that participants had great difficulties relating the correct, 3-frame, small multiple to the animation stimulus that they had seen.

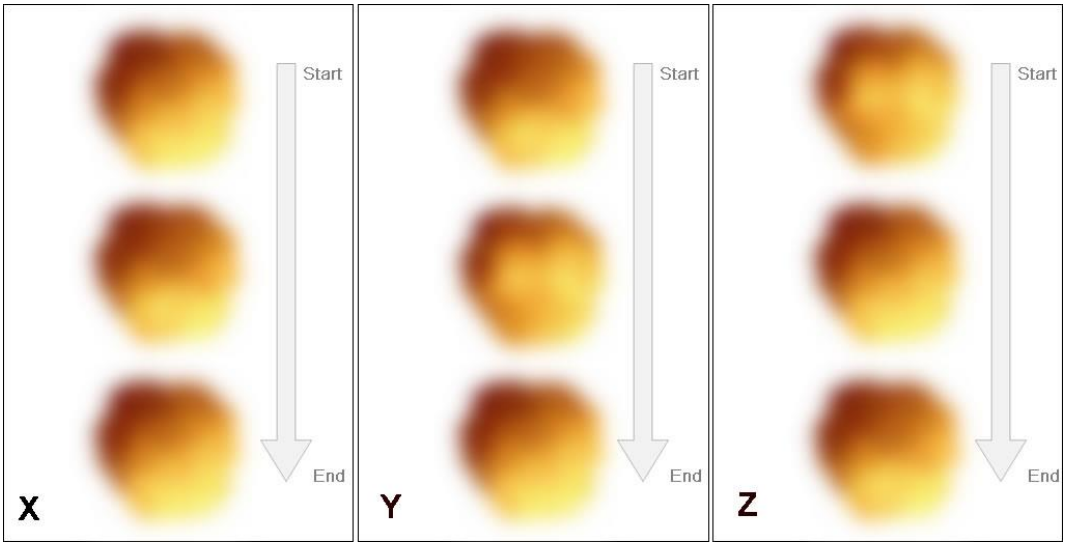

Figure 5: Rejected attempt using static response items for trend detection. Participants in this pre-test had to choose the correct item out of three trend candidates (X, Y, Z)

Thus, we decided to keep presentation mode and temporal granularity, and produced animated response items from downscaled and blurred versions of the actual stimuli and unused test stimulus candidates. While rescaling was necessary to fit response items beside each other on the screen, blurring removed local detail while preserving the overall trend. To ensure that the correct response items were actually determined by their changing global pattern rather than by visually salient outliers, we first removed the two local outliers by assigning them the mean colour of their neighbourhood. Then we downscaled the animations to 160 x 160px and applied a 15px blur (lowpass) filter. Three such global trend response animations (one of which was derived from the actual stimulus animation) were randomly placed next to each other (Figure 6). Each is started by a mouse click and could be replayed as often as desired.

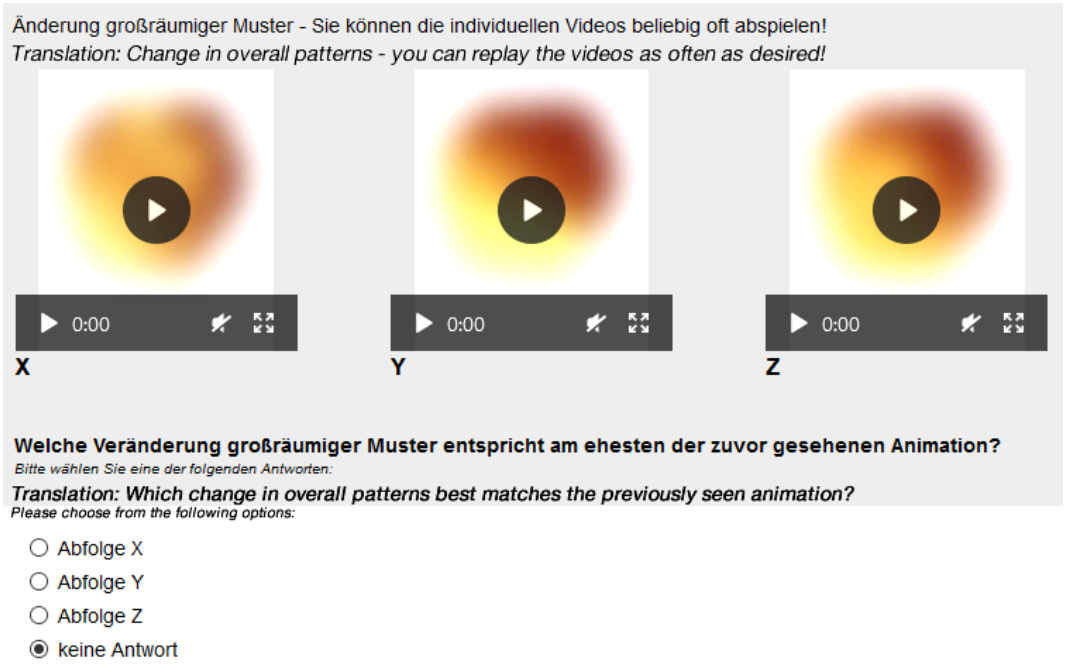

Figure 6: Global trend response item animations. (Original response items did not include English translations.)

3.3 Study design

Despite losing the more closely controlled environment of a laboratory, we chose an online study mode in anticipation of a higher sample size, especially under the current COVID-19 restrictions. We primarily used a four-group between-subject testing design, with the addition of within-subject testing of one stimulus. We did this for two reasons:

1. The anticipated higher power of a within-subject design can act as a 'backup' in case of a very small effect size of generalization, not detectable by between-subject testing.
2. If an effect is detected by the between-subject design (which was the case), we planned to empirically compare both design options in terms of their statistical power by using a repeated subsampling approach.

The results of an extensive pilot study led us to assume that generalization in space might have the largest effects on perception. Therefore, a spatially generalized version and the non-generalized reference of Stimulus 1 were shown to each participant in both parts of the experiment. While half of the participants saw the spatially generalized version first, the other half started with the reference version. To reduce the chance of carry-over effects, these stimulus variants were placed at the beginning (Stimulus 1.1) and the end (Stimulus 1.2) of each of the two test sequences, rotated differently and flipped in a balanced fashion (see Figure 7). In case of (unlikely) carry-over effects, a between-subject comparison of Stimulus 1.1 provides another backup level, thanks to a doubled group size.

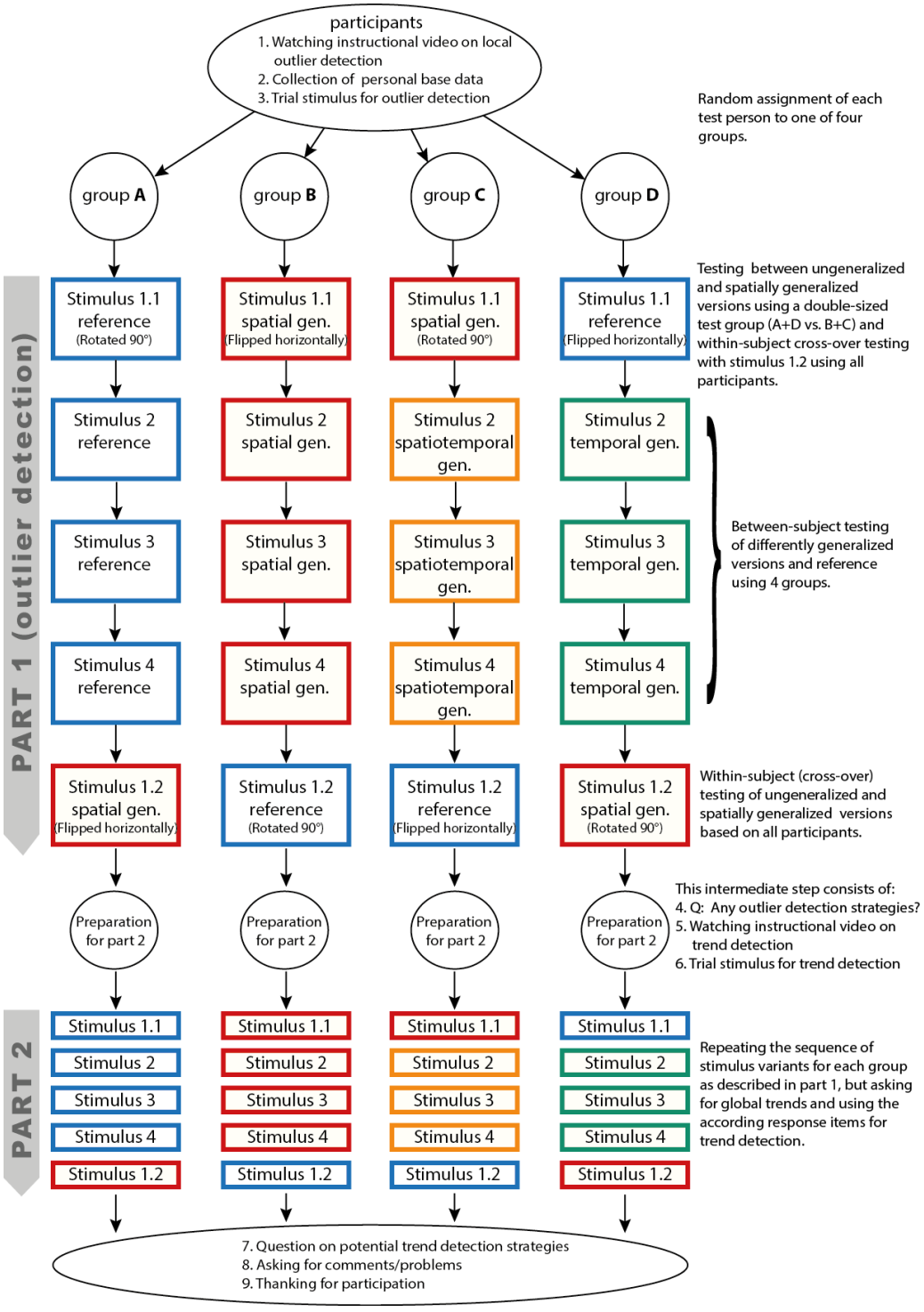

Figure 7: Study design. Source: Traun et al. (2021)

The experiment was implemented as an online study on our university's limesurvey (www.limesurvey.org) instance. To avoid tiny stimuli on small screens and to prevent distraction during participation, we prevented the experiment from running on mobile operating systems and clearly communicated this fact. Forced preloading of stimulus animations in the web-client ensured smooth playback, even in cases of slow internet connection. (Please refer to Figure 7 for the detailed steps in the study procedure.)

3.4 Participants and data

The study was conducted in December 2020 and led – after data cleaning – to a total of 440 responses for analysis, which were distributed equally between the four groups and both genders. Thus, each group (A, B, C, D) provided data from 55 male and 55 female respondents with similar age distributions. (See Traun et al. (2021) for further details on the recruitment of participants and characteristics of the sample.) As in most online studies, participation was motivated primarily by curiosity and personal interest. Such self-selected participation requires critical evaluation as it can bias study results. In our case, we did not see any problem, for two reasons:

1. The study is not related to any personal opinion, but tests basal perceptual abilities. In contrast, participation in a study on personal impacts of the covid-19 crisis might be more attractive to more affected persons. Thus, the aggregated results would potentially be biased by predominant opinions and experiences of a non-representative sample.
2. We assessed potentially confounding factors like cartographic competence, educational level and computer gaming affinity; we did not find any significant effect of these variables on the ability to detect local outliers and global trends.

While analysis and discussion of the data from a cognitive perspective are outside the scope of this paper (see Traun et al. (2021)), the data allowed us to empirically analyse and compare the statistical power of the within-subject and the between-subject designs, as outlined in the following section.

4 Comparing within-subject and between-subject testing

Thanks to our mixed study design, we were able to directly compare both design approaches. For this purpose, we used data on local outlier detection from Stimulus 1. As can be seen in Figure 7, each of the 440 participants saw both variants (ungeneralized reference and spatial generalization) of this stimulus. Half of them (groups A + D) saw the reference version first, while the other half (groups B + C) first saw the spatially generalized version. For each stimulus instance, the number of correctly detected outliers (0, 1 or 2) was recorded.

4.1 Within-subject testing

In the within-subject testing approach, for each participant we calculated the difference in the number of correctly detected outliers between the reference and the spatially generalized versions. For the resulting distribution (Table 2), the 95%-bootstrap-confidence interval (Efron, 1992) for the mean is given by [-0.640, -0.493] and the null hypothesis (mean is equal to 0/there is no difference between spatially generalized and reference variants) is rejected.

Table 2: Differences in correctly identified outliers between reference and spatially generalized animations of Stimulus 1.

Reference – Spatial Variant # of correct outliers	-2	-1	0	1	2
Frequency	42	206	154	36	2

4.2 Between-subject testing

For between-subject testing, we pooled the data from groups A + D (reference in Stimulus 1.1, 220 persons) and tested them against the pooled data from B + C (spatially generalized variant in Stimulus 1.1, 220 persons) using the nonparametric ANOVA-type test statistic from the *R* package *npmv* (see Ellis, Burchett, Harrar, & Bathke, 2017 for further details). The difference is highly significant ($p < 0.001$) and the null hypothesis is rejected. A repetition of this approach using the second instance (Stimulus 1.2) yields the same result. In this type of nonparametric analysis, effect size can be expressed as the probability of achieving a higher response (here: detect more outliers) when belonging to a certain group. With effect sizes of 0.24 (reference) and 0.76 (spatial generalization) for Stimulus 1.1, and of 0.32 and 0.68 for Stimulus 1.2 respectively, the effect of generalization is rather large.

4.3 Comparative analysis

Both design approaches provide evidence that there is a significant difference in outlier detection between the spatially generalized and ungeneralized versions of Stimulus 1 when using our large sample of 440 respondents. But what might have happened if there had been fewer participants (or smaller effect sizes)?

To answer this question, we used repeated subsampling and testing. From each of the eight instances (Stim1.1/group A, Stim1.1/B, Stim1.1/C, Stim1.1/D; Stim1.2/A, Stim1.2/B, Stim1.2/C, Stim1.2/D), we randomly drew the same number of persons while ensuring that nobody was selected twice. This led to subsample sizes of multiples of eight. The within-subject testing was done using the whole subsample, while for the between-subject testing we used only the personal results from the instance a person was drawn from: for example, the correct number of outliers seen in Stim 1.2 group B, if the person was drawn from that instance. These results were pooled into the corresponding spatially generalized group (Stim1.1/B, Stim1.1/C, Stim1.2/A, Stim1.2/D) and reference group (Stim1.1/A, Stim1.1/D, Stim1.2/B, Stim1.2/C), and tested in a between-subject fashion. Note: the inclusion of the second instance of Stimulus 1 was necessary to obtain an unbiased subsample in terms of

evolving learning strategies, as these might play a role in the within-subject case. This subsampling and testing procedure was repeated 500 times for each subsample size (n), each time recording whether the null hypothesis was rejected or not at a given alpha level (we tested at $\alpha = 0.05$ and $\alpha = 0.01$) for both design/testing approaches. The results (Figure 8) show large differences between the within-subject and the 2-group between-subject designs with regard to the total number of participants needed in order to obtain reliable results.

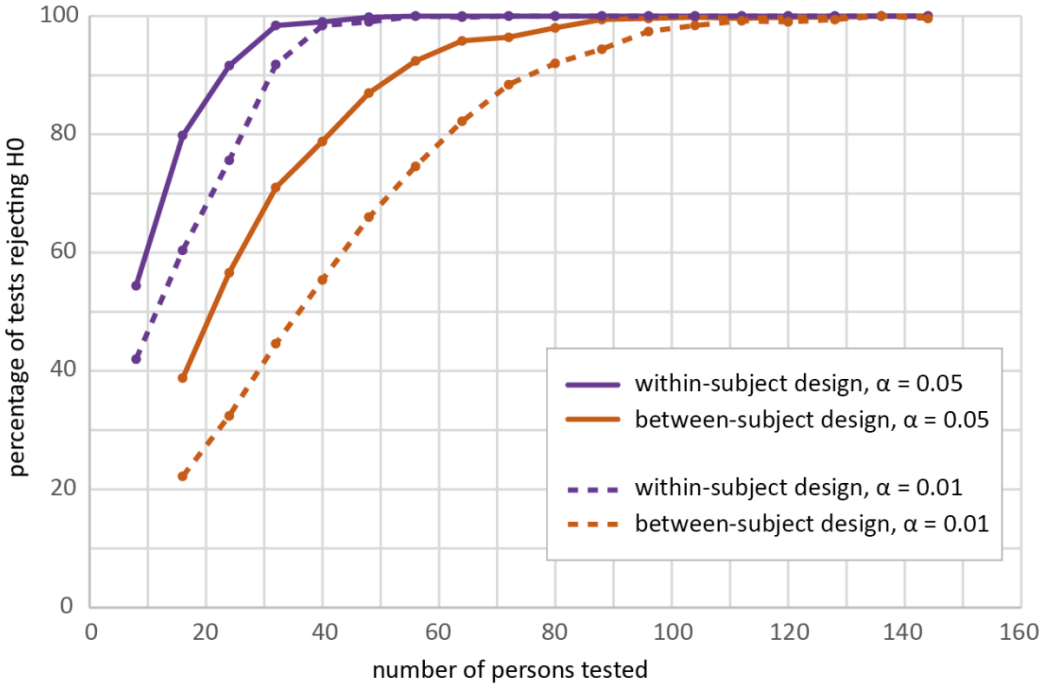

Figure 8: Percentage of the correct detection of significant differences between the spatially generalized and the ungeneralized variant of Stimulus 1 for different sample sizes/numbers of persons tested when using typical within-subject or between-subject approaches. Each data point is based on 500 subsamples/tests. For the between-subject design, 16 individuals (8 per stimulus variant) is the minimum number to achieve test results; for within-subject testing, 8 participants (each generating two data elements to compare) are sufficient for calculations.

Assuming that there is a difference between the spatially generalized and the reference variants, one can see from the simulation results that for an alpha level of 0.05, sample sizes of at least 24 respondents (within-subject testing) and 56 (between-subject testing) are needed for this particular stimulus to reject H0 in 90% of the cases. When more than two groups are to be compared, the total number of participants needed rises accordingly for the between-subject approach. However, having tested 110 persons per group in the remaining Stimuli 2, 3 and 4, we are safe in this respect, while one or the other empirical studies in this field (Table 1) might have been borderline.

To illustrate the difference between within- and between-subject testing (apart from the obvious increase of total persons needed in the latter, explained by a simple multiplication by

the number of groups), normalized results are shown in Figure 9. The total number of test subjects along the x-axis is replaced by the number of (compared) data elements, whereas each within-subject test person is represented by two data elements (Stim 1.1 and 1.2) and each between-subject test person provides just one data element.

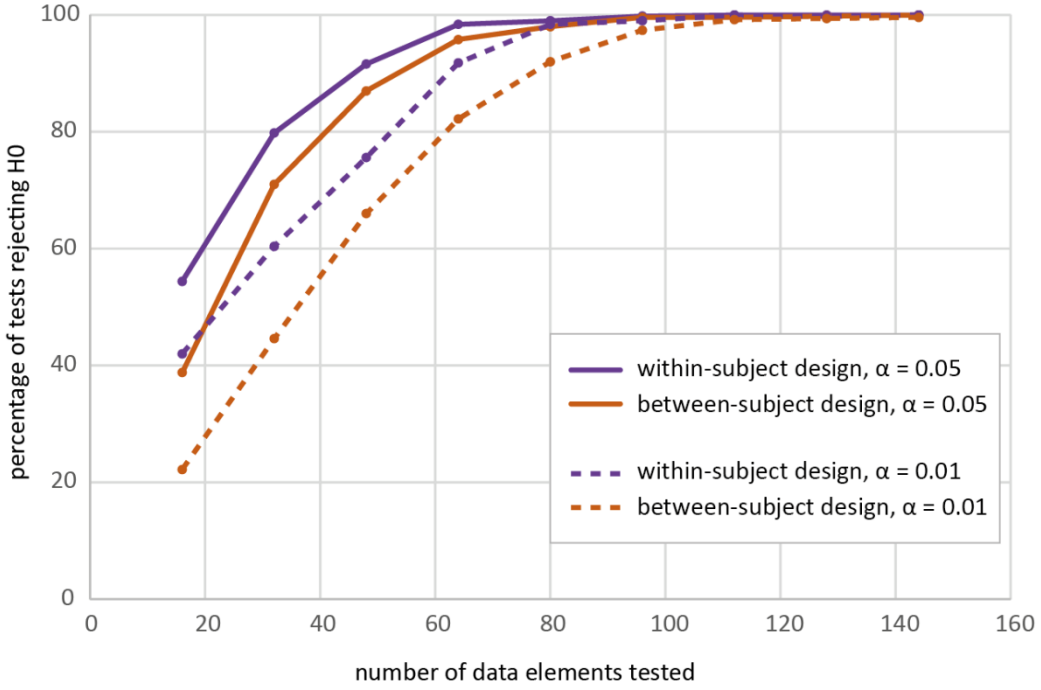

Figure 9: Percentage of tests rejecting H0 by comparing the number of data elements.

The smaller but still systematic difference can foremost be attributed to a better control of confounding variables, such as individual differences in visual perception, motivational level, screen resolution and contrast etc. in the within-subject design, as those variables are largely kept constant. However, the slightly different statistical power of the (nonparametric) bootstrapping approach and the nonparametric ANOVA-type test might also contribute to a difference in either direction.

5 Lessons learned and conclusion

Reviewing the literature, doing pre-testing and running an extensive pilot study provided valuable insights that contributed to the final study design presented here. To give an example, combining outlier and trend detection tasks while viewing the stimuli (although the stimuli were repeated) proved to be a weak point of an early conception used in the pilot study. Having to report on both tasks after the stimulus disappeared added confounding cognitive variables, like divided attention or short-term memory abilities, which diluted data on actual task

performance. Thus, a clear split of the study into two parts was necessary. Although participants still need to recall the outliers (part 1) and trends (part 2) seen in the chosen study setup, we assume the loss of perceived information due to memory imperfections to be invariant under different conditions of generalization, and so not affecting the general perceptive differences we wanted to explore. Study designs allowing for a more direct assessment of perception (like eye tracking) or directly indicating the appearance of an outlier (e.g. verbally or by pressing a button) are superior in this respect, as they are not affected by imperfect recall of information. In our case, however, the cognitive bottleneck is clearly in the perception, as remembering two outlier positions for a few seconds is certainly not a demanding task.

With this paper, we hope to emphasize the diversity of considerations associated with empirical cartographic experiments, notably in the domain of animation, as time adds a level of complexity. The careful design of map stimuli and study architecture play a pivotal role when robust results and transferable insights are the overall goal. Under the advantageous condition of a large effect size in the data we used for power analysis, around 30 participants in a within-subject and around 60 participants in a two-group between-subject design are the minimum to be relatively (around 95%) safe in avoiding type-II errors at the usual alpha level of .05. For smaller effect sizes, higher demands in type-II error prevention and/or smaller alpha levels, larger samples are needed. Apart from cartographic stimulus design in the narrow sense, good question design (Olson, 2009) and counterbalancing (A. Griffin, 2015) are also worth critical consideration when designing cartographic experiments.

Acknowledgements

The first author would like to thank Prof. Dr. Wolfgang Trutschnig for discussing the study design, as well as the anonymous reviewers for their valuable inputs.

References

Battersby, S. E., & Goldsberry, K. P. (2010). Considerations in Design of Transition Behaviors for Dynamic Thematic Maps. Cartographic Perspectives(65), 16-32.

Butler, P. (2016). Mapping Temporal Datasets with D3. Cartographic Perspectives(81), 44-48.

Campbell, C. S., & Egbert, S. L. (1990). Animated cartography/Thirty years of scratching the surface. Cartographica: The International Journal for Geographic Information and Geovisualization, 27(2), 24-46.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. Journal of Economic Behavior & Organization, 81(1), 1-8. doi:https://doi.org/10.1016/j.jebo.2011.08.009

Cybulski, P., & Krassanakis, V. (2021). The Role of the Magnitude of Change in Detecting Fixed Enumeration Units on Dynamic Choropleth Maps. The Cartographic Journal, 1-17. doi:10.1080/00087041.2020.1842146

Cybulski, P., & Medyńska-Gulij, B. (2018). Cartographic redundancy in reducing change blindness in detecting extreme values in spatio-temporal maps. ISPRS International Journal of Geo-Information, 7(1), 8.

DuBois, M. (2013). Complexity and Salience: Evaluating the Inter-Scene Variability of Animated Choropleth Maps. (Master Thesis), University of South Carolina,

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In Breakthroughs in statistics (pp. 569-593): Springer.

Ellis, A. R., Burchett, W. W., Harrar, S. W., & Bathke, A. C. (2017). Nonparametric inference for multivariate data: the R package npmv. Journal of Statistical Software, 76(4), 1-18.

Fish, C., Goldsberry, K. P., & Battersby, S. (2011). Change blindness in animated choropleth maps: an empirical study. Cartography and Geographic Information Science, 38(4), 350-362.

Griffin, A. (2015). Designing your user study or experiment. Paper presented at the ICC 2015, Curitiba, Brazil.

Griffin, A. L., MacEachren, A. M., Hardisty, F., Steiner, E., & Li, B. (2006). A comparison of animated maps with static small-multiple maps for visually identifying space-time clusters. Annals of the Association of American Geographers, 96(4), 740-753.

Harrower, M. (2003). Tips for designing effective animated maps. Cartographic Perspectives(44), 63-65.

Harrower, M. (2007a). The Cognitive Limits of Animated Maps. Cartographica, 42(4), 349-357. doi:10.3138/carto.42.4.349

Harrower, M. (2007b). Unclassed animated choropleth maps. The Cartographic Journal, 44(4), 313-320.

Harrower, M., & Fabrikant, S. I. (2008). The role of map animation for geographic visualization. In M. Dodge, M. M. Derby, & M. Turner (Eds.), Geographic Visualization. Concepts, Tools and Applications (pp. 49-65). Chichester.

Keren, G. (2014). Between-or within-subjects design: A methodological dilemma. A Handbook for Data Analysis in the Behaviorial Sciences, 1, 257-272.

Lobben, A. (2008). Influence of data properties on animated maps. Annals of the Association of American Geographers, 98(3), 583-603.

McCabe, C. A. (2009). Effects of Data Complexity and Map Abstraction on the Perception of Patterns in Infectious Disease Animations. (Master of Science Master Thesis), The Pennsylvania State University, University Park, Pennsylvania.

Monmonier, M. (1996). Temporal generalization for dynamic maps. Cartography and Geographic Information Systems, 23(2), 96-98.

Moon, S., Kim, E.-K., & Hwang, C.-S. (2014). Effects of Spatial Distribution on Change Detection in Animated Choropleth Maps. The Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography, 32(6), 571-580.

Moran, P. A. P. (1950). Note on Continuous Stochastic Phenomena. Biometrika, 37(1), 17-23.

Multimäki, S. (2016). Reducing the information load in map animations as a tool for exploratory analysis. (Dissertation), Aalto University,

Multimäki, S., & Ahonen-Rainio, P. (2015). Temporally Transformed Map Animation for Visual Data Analysis. GEOProcessing 2015, 34.

Olson, J. M. (2009). Issues in human subject testing in cartography and GIS. Paper presented at the Proceedings of the International Cartographic Conference 2009.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? Journal of experimental psychology: Human perception and performance, 18(3), 849.

Traun, C., & Mayrhofer, C. (2018). Complexity reduction in choropleth map animations by autocorrelation weighted generalization of time-series data. Cartography and Geographic Information Science, 45(3), 221-237. doi:10.1080/15230406.2017.1308836

Traun, C., Schreyer, M. L., & Wallentin, G. (2021). Empirical Insights from a Study on Outlier Preserving Value Generalization in Animated Choropleth Maps. ISPRS International Journal of Geo-Information, 10(4), 208.