Chapter 7
A Survey of Variables Used in Empirical Studies for Visualization

ALFIE ABDUL-RAHMAN[1], *King's College London, UK*
MIN CHEN, *University of Oxford, UK*
DAVID H. LAIDLAW, *Brown University, USA*

Abstract
This chapter provides an overview of the variables that have been considered in the controlled and semi-controlled experiments for studying phenomena in visualization. As all controlled and semi-controlled experiments have explicitly defined independent variables, dependent variables, extraneous variables, and operational variables, a survey of these variables allow us to gain a broad prospect of a major aspect of the design space for empirical studies in visualization.

7.1 An Overview of Empirical Studies in Visualization

Empirical studies are an integral part of the research activities in visualization, in a recent survey by Kijmongkolchai et al. [22], some 80 papers on empirical studies, which were published in visualization journals and conferences, were categorized. This is the largest collection to date of papers reviewing controlled empirical studies in visualization, though there are no doubt many more in the literature to be discovered. Many of these empirical studies have provided verifiable means for evaluating different visual designs and visualization techniques, and many others focused on controlled experiments designed to gain some understanding or measurement about specific phenomena in visualization, such as color perception, the effect of emotion, or the use of knowledge.

All controlled empirical studies are designed to study the impacts of the variations of a number of conditions. Mathematically, the individual aspects of the conditions that are being changed during an experiment are defined as *independent variables*, while the effects to be measured are defined as *dependent variables*. Meanwhile, because the variation of an effect could potentially be caused by many variables, each experiment usually has to minimize the impact of some potential variables in order to maintain the total number of conditions being studied at such

[1] e-mail: alfie.abdulrahman@kcl.ac.uk

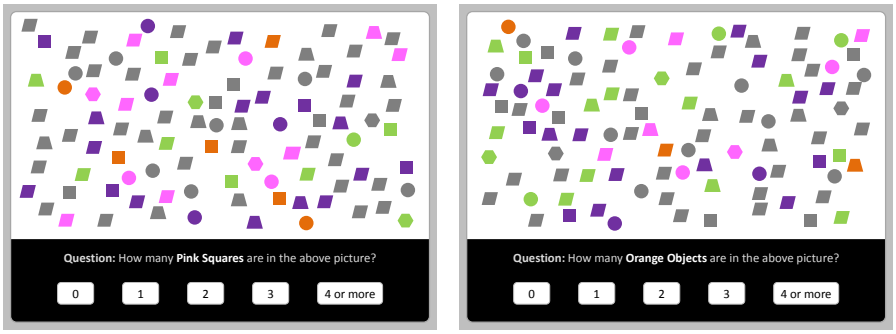
a level that all conditions can be sampled adequately. The methods for controlling a variable other than the pre-defined independent and dependent variables typically include (a) setting it to a constant (e.g., using the same room), and (b) making its instances reasonably random (e.g., ordering different conditions randomly). In some empirical studies, such as web-based crowd sourcing studies, there are well-defined independent and dependent variables, but the impact of some potential variables cannot be fully controlled (e.g., the computer or the room used for the study). They are commonly referred to as *semi-controlled studies*.

There are many forms of empirical studies that do not pre-define a set of independent and dependent variables, including free-text questionnaires, observation diaries, focus group discussions, think aloud sessions, interviews, and so on. One of the goals of such a study is to identify, in an open-minded manner, some independent and dependent variables that may offer potentially the most meaningful explanation about a causal relation in visualization.

In this chapter, we survey the independent and dependent variables that have been studied in controlled and semi-controlled empirical studies in the visualization literature, while examining how extraneous variables were controlled in three case studies. In the remainder of this chapter, we will first give more precise definitions of the main categories of variables. This is followed by a collection of examples for each category. We then detail how variables are defined in three case studies. We offer our summary observation and concluding remarks at the end of the chapter.

7.2 Independent, Dependent, and Other Variables

A *variable* V in an empirical study is a conceptual entity that may change during an experiment, and such an entity can be a piece of stimulus information, a characteristic attribute, an experimental condition, a measurement, or other entity that may vary. For example, in a basic visual search experiment, participants may be



(a) How many pink squares $c_{\text{pink}} \wedge s_{\text{square}}$?

(b) How many orange objects $c_{\text{orange}} \wedge (s_i \in \mathbb{S})$?

Fig. 54 Two example stimuli that may be used for a traditional experiment on visual search.

shown stimuli similar to the two shown in Fig. 54. There are many variables that may change during an experiment, such as:

a. the color of an object in a stimulus;
b. the shape of an object in a stimulus;
c. the size of an object in a stimulus;
d. the position of an object in a stimulus;
e. the number of objects in a stimulus;
f. the aspect ratio of the display area of a stimulus;
g. the questions that may appear in conjunction with a stimulus;
h. the ways in which a question may be answered (e.g., multiple choice buttons, pull down menu, free text, etc.);
i. the number of options available for a multiple choice answer;
j. the ordering of the options available for a multiple choice answer;
k. the type of computing devices used for the experiment;
l. the venue where the experiment is conducted;
m. the time of day when the experiment is conducted;
n. the gender of a participant;
o. the age of a participant;
p. the visual capabilities of a participant;
q. the education background of a participant;
r. the knowledge of a participant that may be used to complete a trial in the experiment;
s. the time taken by a participant to complete a trial in the experiment;
t. the correctness of a participant's response to a stimulus in a trial;
u. the average time taken by all or a specific group of participants to complete the same type of trials in the experiment;
v. the average accuracy of the responses given by all participants or a specific group of participants to the same types of stimuli in the experiment;
w. ...

While we are almost running out of the letters in the English alphabet, it is not difficult to add to the above list. Some variables can be further decomposed into simpler variables. For example, the sight variable may be decomposed to elementary variables of shortsightedness, colorblindness, etc., and the education background may be decomposed to elementary variables of levels, subjects, language, etc.

An empirical study is usually designed to evaluate one or a few hypotheses. Each trial in the experiment is a process that instantiates a causal relation from a set of variables to another set of variables. As a tradition of empirical studies, such a causal relation is usually expressed negatively as a *null hypothesis*. Let a null hypothesis be defined as follows:

> **Null Hypothesis:** Varying variables of $X_1, X_2, \ldots, X_m$ will not have impact on variables of $Y_1, Y_2, \ldots, Y_n$ $(m > 0, n > 0)$.

In general, it is very difficult for an empirical study to evaluate a hypothesis that depends on many variables. Consider, for example, the two stimuli shown in

Fig. 54. There are 100 objects in each stimulus, and each object may have one of the five colors (i.e., c_{green}, c_{grey}, c_{orange}, c_{pink}, and c_{purple}) and in one of the five shapes (i.e., s_{circle}, s_{hexagon}, $s_{\text{parallelogram}}$, s_{square}, and $s_{\text{trapezoid}}$). Hence, each object may appear in one of the 25 color-shape combinations. When considering the 100 objects collectively as a group, the group of objects may appear in 25^{100} color-shape combinations. In other words, if the 100 objects were placed on a fixed regular grid (e.g., a 10×10 grid), there would be 25^{100} possible stimuli. When one takes other variations into consideration, such as the number of the objects, the size of the objects, and the positions of the objects, and so on, the number of possible stimuli will increase rapidly. Since an empirical study can only have a limited number of trials, only a limited number of stimuli can be selected from a vast number of all possible stimuli.

The word "controlled" thus plays a vital role in designing every collected empirical study. Firstly, one has to select a small number of variables of $X_1, X_2, \ldots, X_m$ in a hypothesis by controlling m such that it is a relatively small number (typically $m < 5$). These selected variables are commonly referred to as *independent variables* in the literature of empirical studies. Those variables, which may or may not have an impact on the participants' performance but are not included in the set of, are referred to as *extraneous variables* [7, 21], *nuisance variables* [21], or *potential confounding variables* [13]. An extraneous variable becomes reprehensible as an actual confounding variable, when it is known to have a *confounding effect* on the participants' performance but has not been adequately controlled.

Secondly, one has to control the number of variations or optional values that each independent variable can have. For example, although there are many different colors and shapes that could be used in designing the stimuli in Fig. 54, one has to exercise some control to restrict the number of colors and the number of shapes. To manifest the limited sampling of an independent variable X, one may consider it as an *alphabet* $\mathbb{X}$, which is a term for variable in information theory. The limited number of variations in a variable X is thus the number of letters in the corresponding alphabet $\mathbb{X}$. Therefore, for the experiment illustrated in Fig. 54, the alphabet for the sampled colors $\mathbb{C}$ has five letters, i.e., $\mathbb{C} = \{c_{\text{green}}, c_{\text{grey}}, c_{\text{orange}}, c_{\text{pink}}, c_{\text{purple}}\}$. The alphabet for the sampled shapes $\mathbb{S}$ also has five letters, i.e., $\mathbb{S} = \{s_{\text{circle}}, s_{\text{hexagon}}, s_{\text{parallelogram}}, s_{\text{square}}, s_{\text{trapezoid}}\}$, assuming that the stimuli used in all trials feature the same alphabets $\mathbb{C}$ and $\mathbb{S}$.

Thirdly, one has to control the impact of the extraneous variables, typically by setting each of them to a constant. For example, in the case of the experiment illustrated in Fig. 54, the number of objects is fixed to 100 for all stimuli, and the filled areas of all objects are fixed to the same size. However, not all variables can be fixed to some constants. In many empirical studies, some variables may be sampled randomly, or may appear to be sampled randomly (commonly referred to as pseudo-randomly). For example, the 100 objects in Fig. 54 may appear to be placed in the display area randomly. In fact, they are positioned pseudo-randomly to manifest a reasonably uniform distribution of the objects while avoiding any overlapping, because varying the spatial distribution of the objects would introduce another in-

dependent variable, while varying the amount of occluded part area of each object would undermine the aforementioned control of the object size.

Finally, many variables can neither be fixed to constants nor be sampled randomly or pseudo-randomly. For example, it would be very difficult to fix the ages and education backgrounds of the participants to some constants, or to recruit participants in a way reflecting a uniform distribution. In such cases, the common wisdom is to record the variations of such variables in an empirical study and discuss their potential impact in the report of the experiment. In some situations, one may determine whether or not such a variable has an impact on the hypothesized causal relation. In most other cases, one may have to leave such conclusions to some future empirical studies.

The set of variables of $Y_1, Y_2, \ldots, Y_n$ in the general formation of a null hypothesis are referred to as *dependent variables*. There are two main classes of dependent variables. The variables that are to be measured in individual trials are *measured dependent variables*. The most elementary dependent variable is a binary variable. Most experiments for studying just notifiable difference (JND) ask participants to choose whether the attribute of one stimulus is above or below that of another stimulus. Many experiments for testing rapid reaction or decision capacities also use binary variable, such as "yes" or "no", "on" or "off", "action" or "no action", and so on.

The slightly more complicated dependent variable is a set of multiple choices, typically implemented as multiple command buttons, radio buttons, or selectable visual objects in a stimulus. The examples in Fig. 54 show five command buttons. Hence, when the multiple choices are considered as letters of an alphabet, we have an alphabet for the answer $\mathbb{A} = \{a_0, a_1, a_2, a_3, a_{4+}\}$.

Some empirical studies have much more complicated alphabets as measured dependent variables. For example, selecting a location on a map from n optional locations or entering a real number with high precision involves a very large alphabet. Later in Section 7.4.3, we will see an empirical study that captures 14 time series as measured dependent variables.

From one or more measured variables, one may define a *derived dependent variable*. For example, one may define the correct answer of Fig. 54(a) is a_0 and that of Fig. 54(b) is a_3. With such defined ground truth information, one may define the *correctness* of each trial as a *derived dependent variable*. By aggregating the correctness values of a group of trials, one can define accuracy (in percentage) as a derived dependent variable for the group. Similarly, the response time of a participant in each individual trial is a measured dependent variable, while the mean response time for a group of trials is a derived dependent variable. The way in which the trials are grouped together depends on the hypothesis concerned, the definition of independent variables, and the control of extraneous variables.

In order to compute a derived dependent variable from a measured variable, one has to use some additional definitions (e.g., the ground truth) and additional functions (e.g., statistical or algorithmic functions). The variation of such a definition or a function would have an impact on the derived dependent variable concerned.

Hence, these definitions and functions are also variables, which are referred to as *operational variables* or *operational definitions* [21].

7.3 Examples of Variables Used in Empirical Studies

In this section, we first provide three lists of typical variables resulting from our surveys of the papers collected by Kijmongkolchai et al. in [22]. In particular, we conducted close-reading of 32 papers that report controlled and semi-controlled empirical studies in visualization, and identified all variables in these papers. In Section 7.4, we will detail our analysis of the variables in three examples of empirical studies, which represent quite different study designs.

7.3.1 Independent Variables

There are numerous independent variables that have been studied in different empirical studies. It is not feasible to list all these variables exhaustively. Following a careful reading of 32 papers on visualization-related empirical studies, we identified some 50 variables, and categorize them into five classes.

7.3.1.1 Varying Values in a Single Visual Channel or Varying Types of Visual Channels.

The first class is *elementary visual channels* (or *elementary visual variables*), which have been often featured in studies that investigate the attentiveness, distinguishable values, and metaphoric association of different visual channels, as well as the differentiation and interaction between them. The following list gives a number of examples used in several empirical studies. Each item listed, X, can be read as "varying X in the stimuli."

We note that many empirical studies feature stimuli with different visual channels. When a study was not designed to evaluate any hypothesis suggesting that varying such visual channels might have an impact, we do not consider them independent variables of the study. For example, Szafir [32] conducted an empirical study to investigate whether varying the size of graphical primitives impacts color perception. There were extraneous variables associated with the graphical primitives, which were not part of the hypotheses. A polyline primitive, for instance, features many data points, which are extraneous variables that determine the shape of the polyline. The study focused on the thickness of polylines as an independent variable, while controlling other extraneous variables such as the overall height and width, the number of data points, and so on.

- color differences (their levels) [32];

- colors (of glyphs) [12];
- shapes (of glyphs) [12];
- sizes (of glyphs) [12];
- sizes (of graphical primitives) [32];
- types of visual channels (for indicating grouping) [2];
- types of visual channels (for values of missing data) [30];
- vector magnitude [36].

7.3.1.2 Varying Visual Objects Featuring Multiple Visual Channels or the Characteristic Attributes of the Combined Variations

This class of independent variables features variations of multiple visual channels of some visual objects in stimuli. The goal of such a study is typically to investigate the interaction or the combined effects of more than one visual channel. In some cases, the experimenters may focus on a single independent variable that characterizes the combined variations of multiple visual channels, such as the ordering of colors in a colormap [29]. Because the variation of the ordering in this case is more complicated than the variation of a single color, we consider such an independent variable falls into this class.

- bi-variate channels (the shape-color combinations) [12];
- bi-variate channels (the shape-size combinations) [12];
- bi-variate channels (the size-color combinations) [12];
- continuous colormaps (their key colors) [4];
- continuous colormaps (their ordering of key colors) [29];
- discrete colormaps (palette sizes) [14];
- discrete colormaps (palette scoring functions) [14];
- discrete colormaps (user-generated vs. software recommended vs. random) [14];
- discrete colormaps (with semantic association or not) [29];
- multivariate channels (the combinations of 2-5 channels used for indicating grouping); [2];
- multivariate channels (for map textures) [24].

7.3.1.3 Varying Visual Patterns Made of Multiple Visual Objects or the Characteristic Attributes of the Visual Patterns

This class features variations of what one common referred to as “patterns”. A pattern is considered to be made of multiple visual objects. Typical examples include a cluster in a scatter plot or dot plots, an ego or focal node in a network visualization, a volatile section in a time series plot, etc. In general, the variation of patterns involves the simultaneous variations of several visual objects, and is thus considered to be more complicated than the variation of a few visual channels of the same visual object as discussed in Section 7.3.1.2.

Because the possible number of such variations is usually excessively large and their distribution in a context is often not well established, it is difficult to create a set of stimuli that constitute an unbiased sampling of the space of such variations. It is thus common to control the sampling by introducing some characteristic attributes (e.g., levels of complexity or sparseness, types of ordering or configuration, and so on), and making such attributes as the independent variables.

- data characteristics (level of deviation from a trend-line) [9];
- data characteristics (densities) [18];
- data characteristics (gap, flow-type outlier, spike) [10];
- data characteristics (levels of noise) [28];
- data characteristics (trend types) [9];
- feature patterns (in dot plots) [27];
- feature patterns (ordering of visual objects) [37];
- feature patterns (simple vs. complex) [22];
- highlighting methods (color, leader line) [16];
- levels of appearance fidelity (of virtual human avatars) [34];
- levels of negative emotions (time-steps) [34];
- pixel patterns (block resolutions) [4];
- pixel patterns (block sizes) [15];
- pixel patterns (pixel sizes) [15];
- pixel patterns (subset configurations) [15];
- pixel patterns (levels of variety) [17];
- pixel patterns (types of variety: color or motion) [17];
- pixel patterns (types of variety: local vs. global) [17];
- word-tag patterns (area of words) [11];
- word-tag patterns (colors of word tags) [11];
- word-tag patterns (densities of word tags) [11];
- word-tag patterns (lengths of word tags) [11];
- word-tag patterns (area of words and types of word spacing) [11].

7.3.1.4 Varying Plot Types or Plot-level Visual Designs

The independent variables in this class define variations at the plot level, and are typically used to compare different visual representations or significant variations of visual designs of a type of plots.

- multi-plots (multi-view compositions: map with scatter plot vs. map with parallel coordinates plots) [16];
- plot attributes (aspect ratios) [18];
- plot attributes (chart height and virtual resolution) [19];
- plot attributes (chart height and gridline spacing) [18];
- plot types (nine types of plots) [18];
- plot types (braided graph, horizon graph, line graph, small multiples) [20];
- plot types (density plot, gap-detection histogram, dot plot) [10];

- plot types (graph, scatter plot, storyline, treemap) [33];
- plot types (filled line chart, mirrored chart, 2-band horizon chart) [19];
- plot types (line graph, colorfield) [8];
- plot types (scatter plot, line graph, area) [9];
- visual designs (2D flow visualization) [23];
- visual designs (bar charts and difference overlays) [31];
- visual designs (for map-based flow visualization) [35];
- visual designs (with or without embellishment) [3]).

7.3.1.5 Varying Variables not in the Depicted Data

The effectiveness of visualization does not only depend on the depicted data values, the selection of visual channels or the design of visual representations, but also on many other factors such as user, task, application, and so on. This class thus includes all independent variables that are used to study the impact of such factors on visualization processes.

- display types (mono, stereo) [36];
- display types (MacOS, others) [18];
- teaching methods (bottom-up, top-down) [33];
- application contexts [22];
- color compensation configurations [26];
- statistical measures (min/max, mean, stdev) [22];
- learning approaches (passive, active) [33];
- visualization tasks (many studies, e.g., [23, 4, 18, 3, 31]).

7.3.2 Dependent Variables and Derived Variables

The variables that are to be measured in individual trials are *measured dependent variables*. In most cases, the collected values of some dependent variables are processed to yield some numerical quantities or categorical values, using, e.g., statistics or algorithms, we consider the corresponding variables as *derived dependent variables*.

There are a number of measured dependent variables that commonly defined in many empirical studies, including:

- response time (RT) of a trial;
- a selection out of k choices ($k \geq 2$);
- a value entered using a 1D scroll bar;
- a position in a 2D map entered using a pointer device (often with many optional locations);
- a location in a 3D real or virtual environment entered using a 3D input device;

- a sequence of action records (e.g., user interactions, and navigation actions in a virtual environment);
- an eye-tracking record;
- one or more time series records of EEG (electroencephalography);
- one or more imagery records of fMRI (functional magnetic resonance imaging or functional MRI).

During an empirical study, some measured dependent variables may be used to compute some derived variables dynamically. Perhaps the most common variable derived dynamically is correctness indicated by a measured value in order for the experiment system to give a feedback to the participant. For example, a system for facilitating trials with multiple-choice questions may maintain the ground truth answer for each trial, and use it to determine the correctness of an answer. A system for eye-tracking may maintain a set of areas of interest, and use these to determine if a participant's gaze has been fixated on any of the areas of interest.

Because these derived dependent variables are obtained using some predefined operational variables such as ground truth values, threshold values, quantization bands, etc., they are not only dependent on the input stimuli and the human actions during trials, but also on these operational variables. Hence it is helpful to consider them as derived dependent variables in order to be mindful about the variations of the underlying operational variables and functions that could affect the findings of the study.

In almost all empirical studies, the analysis of the results involve derived dependent variables defined through statistical aggregation and analysis. The most commonly used derived dependent variables are:

- accuracy and error rate (percentage values calculated based on a collection of correctness values);
- precision and recall (for information retrieval tasks);
- just notifiable difference (JND);
- average response time (mean RT, often abbreviated as RT);
- basic statistical measures for a collection of measured or derived values (e.g., mean, max, mean median, mode, range, correlation coefficient, mutual information, etc.);
- measures resulting from processes of statistical analysis, such as t-test, χ^2-test, ANOVA (analysis of variance), and so on.

In experiments designed with some specific apparatus, there are usually some specialized dependent variables. For example, in eye-tracking experiments, one may define (a) time from the start of a trial to the first fixation at an area of interest and (b) the number of fixations during a trial as derived dependent variables computed based on gaze records [16]. A number of studies measured specific types of participants' judgment, such as alpha contrast optimization [18], discriminability rate [32], perceived complexity [28], perceived data quality [30], and so on. Using an electrodermal activity (EDA) sensor, one may obtain an EDA data set as measured dependent variables, and may compute differential emotions scale (DES) as a derived dependent variable [34]. In a recent empirical study, the traditional accuracy

and mean RT variables were transformed to information-theoretic measures of benefit and cost as a new pair of derived dependent variables [22].

In general, determining a collection of variables that may affect a design provides a means for defining a design space. In visualization, some notable publications (e.g., [5]) proposed and discussed design spaces of visual representations. One may wonder if there might be a design space for controlled empirical studies in visualization. Our enumeration of experimental variables here may begin to inform the description of such a space. However, given the level of complexity that arises from just this simple initial step, formulating a structured description of such a design space seems to be out of reach at the moment. We hope that a design space for empirical studies in visualization will emerge in the future.

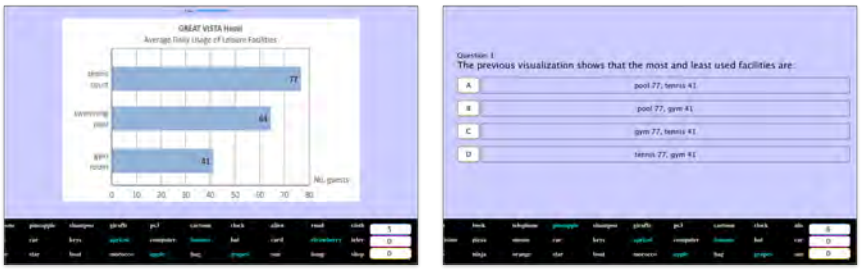
7.4 Case Studies

In this section, we present three case studies to show how one may extract the information of independent, dependent, and constrained variables. In psychology, many papers reporting empirical studies define independent and dependent variables explicitly. In those papers that do not offer explicit definitions of the study variables, it is usually not too difficult to extract such information indirectly. In general, the stimuli used in visualization-related experiments are more complicated, and it is not always easy to extract the definitions about such variables. For the three papers discussed in this section, the authors of this chapter first read the papers and wrote down the independent and dependent variables individually. They then compared the notes, and agreed on a common set of variables.

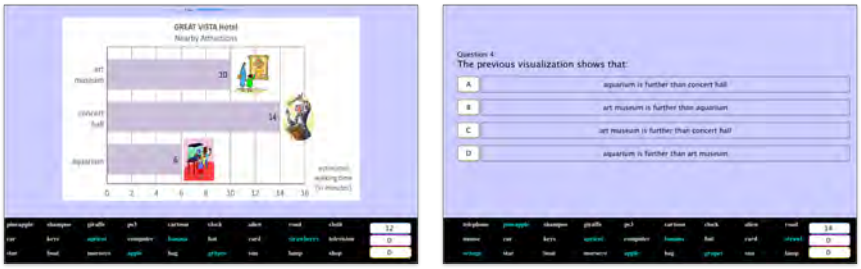
7.4.1 A Study on Using Visual Embellishments in Visualization

Borgo et al. presented a study on the impact of visual embellishment on participants' ability (a) to remember the numerical data depicted, (b) to perform visual search of visual objects, and (c) to grasp the concept conveyed by the text shown in visualization images [3]. Because the tasks in such a study had to be reasonably simple in order to control the potential confounding effects and the length of each trial, they anticipated that the impact might not be easily detectable if the participants were paying attention to their tasks. They thus designed a dual-task experiment, where a secondary task was used to restrict the amount of the cognitive capability available to the primary task in each trial, allowing the trials with embellishment and those without more differentiable.

For the primary tasks, all stimuli were designed in pairs, one with visual embellishment and one without. Hence this binary variable was the most important independent variable being studied. The experimenters had four hypotheses and they divided the stimuli into four sections. Since these four sections were conducted within



(a) A stimulus screen and its follow-on question screen



(b) A related stimulus screen and its follow-on question screen

Fig. 55 Two related trials in the empirical study. The top 80% of each screen was used for the primary task and the bottom 20% was used for the secondary task. A stimulus without any visual embellishment is shown in (a) and a stimulus with a similar visual representation and a similar amount of information as well as with some visual embellishment is shown in (b). The two trials were distributed pseudo-randomly among others to minimize any learning effect.

the same empirical study, the four topics, i.e., working memory, long-term memory, visual search, and concept grasping were the four values of a variable about tasks. Thus there were two independent variables for the primary tasks.

To avoid learning effects, different stimuli had to feature different data values. These are extraneous variables that should be controlled. The experimenters carefully selected these values to ensure a similar level of complexity within each pair of stimuli, while having different levels of complexity across different pairs for each section of the experiment. Both measures provided means of controlling the potential confounding effects due to the variations of the data values.

Similarly there were variations of the designs for different visual embellishment across different stimuli. Such variations were unavoidable since each trial featured a different dataset and it was necessary to change the semantics featured the datasets to avoid learning effect. The experimenters controlled the potential confounding effects due to such an extraneous variable by using the same approach for dealing with the variations of data values.

For the primary task, each trial presented participants with a question and four optional answers (one correct answer and three distractors). Hence the measured dependent variables were the selection out of four options and the time taken to make this selection. By pre-defining the ground truth value of each trial (i.e., an

operational variable), the experimenters obtained a derived dependent variable for the correctness of the selected answer. Similar to numerous empirical studies, from the correctness and response time of each trial, the two commonly-used dependent variables were derived, i.e., the mean accuracy and the mean response time.

The stimuli for the secondary task ran continually in parallel with the stimuli for the primary task throughout the experiment. In a rectangular area at the bottom of the screen, a sequence of words moved horizontally from left to right, with new words appearing from the left continually. Participants were required to point and click at any fruit word that appeared in that area. When each word displayed was considered as a visual object, from the perspective of the secondary task, each word had only two states, a fruit word or a non-fruit word. Thus, the independent variable was a binary variable. When a participant selected a word, the dependent variable was correctness. The software for the experiment showed three counters on the screen, keeping the count of how many fruit words had been correctly selected, how many had been missed, and how many words had been wrongly selected. These counters were derived dependent variables.

In addition to the aforementioned effort for controlling the potential confounding effects due to the variations of data values and visual embellishment, the experimenters also discussed effort for controlling other extraneous variables, such as knowledge bias, ordering bias, and attention bias.

7.4.2 A Study on Visual Semiotics and Uncertainty Visualization

MacEachren et al. presented two controlled empirical studies on aspects of uncertainty visualization [25]. The first experiment was designed to obtain measurements about the participants' judgment as to the suitability of visual representations for a given category of uncertainty. They defined their first independent variable for ten categories of uncertainty, which were referred to as ten series in their paper [25]. The alphabet $\mathbb{X}_{\text{series}}$ thus consists of 10 letters: (x_1) general, (x_2) spatial accuracy, (x_3) spatial prevision, (x_4) spatial trustworthiness, (x_5) temporal accuracy, (x_6) temporal precision, (x_7) temporal trustworthiness, (x_8) attribute accuracy, (x_9) attribute precision, and (x_{10}) attribute trustworthiness.

The letters x_2-x_{10} were defined over two elementary alphabets. One alphabet defined three categories of data to be displayed (i.e., space, time, and attribute), and the other defined three types of uncertainty associated with the data (i.e., accuracy, precision, and trustworthiness). The letters x_2-x_{10} were the nine combinations of the letters of these two elementary alphabets.

The second independent variable $\mathbb{X}_{\text{level}}$ defined the two levels of abstraction of the symbol sets: namely abstract or iconic. Each symbol set consisted of k glyphs that represented different levels of uncertainty. In this experiment, k was considered as an extraneous variable, which was fixed to $k = 3$.

The experimenter designed 76 symbol sets for 76 trials. They were used primarily as repeated measures of the two levels of abstraction. For series x_1, 22 symbol

sets were used, and the symbol sets were designed based on different visual channels (e.g., color, size, shape, etc.). For each of series x_2-x_{10}, six symbol sets (three abstract and three iconic) were used. The variation of symbol sets was a variable difficult to control, because it was not easy to define the design space of the symbol sets. The experimenters made a good effort to design various symbol sets considered to be most representative and sensible designs heuristically. They recorded and reported the impact of individual symbol set on the participants' judgment, exhibiting the best practice for handling such an extraneous variable.

The measured dependent variable was the subjective judgment in each trial by a participant. The corresponding alphabet $\mathbb{Y}_{\text{judgment}}$ consists of seven levels of intuitiveness of a symbol set, i.e., $\mathbb{Y}_{\text{judgment}} = \{1,2,3,4,5,6,7\}$. It was implemented using a set of clickable numbers for the seven multiple choices. From this measured dependent variable, the experimenters computed a set of derived dependent variables, including five statistical measures (i.e., min, max, mean, median, and mode) and two measures of the Mann-Whitney test (i.e., W and p-value).

In addition, the experimenters obtained a measured dependent variable of the time taken to complete each trial. They reported three derived dependent variables resulting from the independent two-group t-test with Welsh df-modification (i.e., t, df, and p-value).

The second experiment was designed to obtain the measurements about the effectiveness of the symbol sets through a typical task in map visualization. Participants were asked to assess and compare the aggregated uncertainty in two map regions based on the glyph representation of uncertainty in each location. The experiment featured one independent variable that defines 20 symbol sets selected based on the results of the first experiment. In other words, it was an alphabet with 20 letters. The goal of the experiment was to determine the relative merits among these 20 symbol sets.

In the context of this experiment, a map being visualized can be considered as a background image, and the uncertainty glyphs can be placed on a $w \times h$ grid superimposed on top of the background map. The variations of the map image and the grid resolution would manifest variables with very large sampling spaces. The experimenters considered them as extraneous variables, and controlled both of them by using constants. The background image was simply removed from all stimuli, while the grid resolution was fixed to 3×3.

The stimulus in each trial depicted two regions, each with 3×3 uncertainty glyphs. All 18 glyphs in each stimulus were selected from the same symbol sets. The task of each participant was to aggregate the nine uncertainty values in each of the two regions, and select the region that was less certain. Since each symbol set had three glyphs representing three uncertainty values, there were a total of 3^{18} possible variations of the stimuli. The experimenters controlled this extraneous variable using pseudo-randomness by pre-defining 12 configurations that represented a relatively uniform sampling of the stimuli space. Although varying the 12 configurations could be considered as an independent variable, they were featured in the experiment design as an extraneous variable for supporting repeated measures for

each symbol set. Together, the experiment had a total of 240 trials (20 symbol sets and 12 configurations).

The measured dependent variables were the correctness of a participant's selection and the response time in each trial. The derived dependent variables reported in [25] included the accuracy of 20 symbol sets, and the accuracy value for each symbol set was an aggregation of the 360 correctness values (the 30 participants and 12 configurations). In addition, the experimenters applied the Pearson's χ^2 test with Yates' continuity correction to the correctness values, yielding three derived dependent variables χ^2, *df*, and *p*-value; and applied the independent two-group *t*-test with Welsh *df*-modification to response time, yielding three derived dependent variables *t*, *df*, and *p*-value.

7.4.3 An EEG Study on Visualization Effectiveness

Anderson et al. presented an empirical study on participants' cognitive load during visualizing different visual designs of box plots [1]. In each trial, a participant was shown two types of box plots with different data and was asked to choose the distribution with a larger inter-quartile range.

The main independent variable defined the variations among six visual designs of box plots. The corresponding alphabet had six letters. Most box plots typically depicted five statistical measures computed over a data sample, including (i) minimum, (ii) median, (iii) maximum, (iv) the 25th percentile, and (v) the 75th percentile. Most visual designs allowed the viewers to estimate the min-max range and the inter-quartile range (between the 25th and 75th percentile). Some box plots also depicted the distribution of data values in the sample using a visual representation based on histogram or a density map. The experimenters selected three visual designs with a density map and three without. This additional independent variable allowed the evaluation of a hypothesis related to the absence/presence of the distribution information.

The summary statistical measures depicted by a box plot were computed from *n* values in a data sample. The variations of the data sample determined the variations of its statistical measures, and hence the corresponding box plot. The data space for *n* values was exponentially related to *n*. The experimenters had to control such variations. In this experiment, the extraneous variable of data samples was controlled firstly by fixing the number of data values to 100 and the distribution of the sample to uniform, and secondly by using randomly generated data values with controlled ranges for the mean and standard division of the sample. A total of 500 samples were generated, hence there was a pool of 500 box plots. In the study, each participant performed tasks in 100 trials, each of which showed two box plots selected from the pool.

Given two samples, the experimenters estimated the task difficulty, in the range of [0, 1], of comparing the two corresponding box plots. This variable was not explicitly featured in the stimuli, but was used in results analysis as a possible cause

that might impact the cognitive load. One could consider this as an independent variable.

For each trial, the experiment captured a number of measured dependent variables, including (a) the response time, and (b) the electroencephalography (EEG) signals in the form of 14 time series. The experimenters used a numerical function for transforming the 14 time series to a derived dependent variable to as the estimated cognitive load per trial. From the estimated values for cognitive load for all trials, two further derived dependent variables were computed, namely constant- and Gaussian-weighed averages. They also applied two-tailed t-tests to compare the cognitive load values estimated for every pair of visual designs, and obtained the 15 p-values for the corresponding derived dependent variables.

7.5 Conclusions

In this chapter, we have conducted a survey on independent and dependent variables used in controlled or semi-controlled empirical studies on the subject of visualization. In particular, we analyzed the variables considered in 32 publications on such studies. We categorized independent variables into five categories. We noticed that there are no shortage of studies on independent variables in each category. We consider this a particularly encouraging sign, because this shows that visualization researchers are asking many research questions about visualization at different levels of visual designs and from many different perspectives. Meanwhile it also suggests that there are many more research questions yet to be asked or answered, and the scope of visualization-related empirical studies is huge.

Meanwhile, when an independent variable is examined in one study, it can be an extraneous variable to be controlled in another study. The variety of independent variables that have already been examined in the previous studies indicate the challenge in alleviating confounding effects since controlling many extraneous variables is not a trivial undertaking in most visualization-related empirical studies

The large number of variables and potential experimental designs also brings up the point that designing experiments is a creative process. As with any process that involves design, there are many choices to be made in many trade-offs that need to be balanced in making those choices. There is no one best design, just as there is no one best painting, building, or software application. Learning to design good experiments is a matter of study and practice, and there are numerous books and other resources that teach how to do it. We have touched on a few of the design decisions and trade-offs that we identified in the visualization literature, but this survey is only a sparse sampling of the rich space of experimental design.

It is hence necessary for the experiment designers to be aware of the potential impact of different extraneous variables is important, while it is helpful for the reviewers to appreciate the challenge of alleviating confounding effects. Occasionally, some of us in the community may wish the stimuli in some empirical studies to be more complex or more realistic without appreciating that more complex or more

realistic stimuli would likely introduce more confounding effects that could undermine the statistical significance of the experiment results. In other occasions, some of us in the community may wish that the stimuli in some empirical studies could feature fewer independent variables or extraneous variables could be controlled more stringently without being aware of the experimenters' intention to examine the impact of variables at a higher level (e.g., multi-object patterns or plot-level visual designs).

It may thus be desirable for the visualization researchers who conduct empirical studies to be more coherently organized, instead of being distributed sparsely in InfoVis, SciVis, VAST, and other areas of visualization. This will allow these researchers to share their expertise (e.g., in the review processes) more easily and to formulate research agenda in a more ambitious and structured manner. If one considers different schools of thought in visualization (see Chapter 11 [6]) as high-level hypotheses, there are indeed many ambitious research questions that may be answered using empirical studies. By providing some opportunities to bring all these researchers together, we may soon see the emergence of a new area of visualization psychology.

References

1. Anderson, E.W., Potter, K.C., Matzen, L.E., Shepherd, J.F., Preston, G.A., Silva, C.T.: A user study of visualization effectiveness using EEG and cognitive load. Computer Graphics Forum **30**(3), 791–800 (2011)
2. Bae, J., Watson, B.: Reinforcing visual grouping cues to communicate complex informational structure. IEEE Trans. Visualization & Computer Graphics **20**(12), 1973–1982 (2014)
3. Borgo, R., Abdul-Rahman, A., Mohamed, F., Grant, P.W., Reppa, I., Floridi, L., Chen, M.: An empirical study on using visual embellishments in visualization. IEEE Transactions on Visualization and Computer Graphics **18**(12), 2759–2768 (2012)
4. Borgo, R., Proctor, K., Chen, M., Jänicke, H., Murray, T., Thornton, I.M.: Evaluating the impact of task demands and block resolution on the effectiveness of pixel-based visualization. IEEE Trans. Visualization & Computer Graphics **16**(6), 963–972 (2010)
5. Card, S.K., Mackinlay, J.: The structure of the information visualization design space. In: Proc. IEEE Symposium on Information Visualization, pp. 92–99 (1997)
6. Chen, M., Edwards, D.J.: Isms in visualization. In: M. Chen, H. Hauser, P. Rheingans, G. Scheuermann (eds.) Foundations of Data Visualization. Springer (2019)
7. Cohen, P.R.: Methods for Artificial Intelligence. MIT Press Cambridge, MA, USA (1995)
8. Correll, M., Albers, D., Franconeri, S., Gleicher, M.: Comparing averages in time series data. In: Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 1095–1104 (2012)
9. Correll, M., Heer, J.: Regression by eye: Estimating trends in bivariate visualizations. In: Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 1387–1396 (2017)
10. Correll, M., Li, M., Kindlmann, G., Scheidegger, C.: Looks good to me: Visualizations as sanity checks. IEEE Trans. Visualization & Computer Graphics **25**(1), 830–839 (2019)
11. Correll, M.A., Alexander, E.C., Gleicher, M.: Quantity estimation in visualizations of tagged text. In: Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 2697–2706 (2013)
12. Demiralp, C., Bernstein, M.S., Heer, J.: Learning perceptual kernels for visualization design. IEEE Trans. Visualization & Computer Graphics **20**(12), 1933–1942 (2014)

13. Eysenck, M.W.: Psychology: A Student's Handbook. Psychology Press (2000)
14. Gramazio, C.C., Laidlaw, D.H., Schloss, K.B.: Colorgorical: Creating discriminable and preferable color palettes for information visualization. IEEE Trans. Visualization & Computer Graphics **23**(1), 521–530 (2017)
15. Gramazio, C.C., Schloss, K.B., Laidlaw, D.H.: The relation between visualization size, grouping, and user performance. IEEE Trans. Visualization & Computer Graphics **20**(12), 1953–1962 (2014)
16. Griffin, A.L., Robinson, A.C.: Comparing color and leader line highlighting strategies in coordinated view geovisualizations. IEEE Trans. Visualization & Computer Graphics **21**(3), 339–349 (2015)
17. Haroz, S., Whitney, D.: How capacity limits of attention influence information visualization effectiveness. IEEE Trans. Visualization & Computer Graphics **18**(12), 2402–2410 (2012)
18. Heer, J., Bostock, M.: Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In: Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 203–212 (2010)
19. Heer, J., Kong, N., Agrawala, M.: Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In: Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 1303–1312 (2009)
20. Javed, W., McDonnel, B., Elmqvist, N.: Graphical perception of multiple time series. IEEE Trans. Visualization & Computer Graphics **16**(6), 927–934 (2010)
21. Kantowitz, B., Roediger III, H., Elmes, D.: Experimental Psychology. Wadsworth Publishing (2014)
22. Kijmongkolchai, N., Abdul-Rahman, A., Chen, M.: Empirically measuring soft knowledge in visualization. Computer Graphics Forum **36**(3), 73–85 (2017). DOI 10.1111/cgf.13169. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13169
23. Laidlaw, D.H., Davidson, J.S., Miller, T.S., da Silva, M., Kirby, R.M., Warren, W.H., Tarr, M.: Quantitative comparative evaluation of 2D vector field visualization methods. In: Proc. IEEE Visualization, pp. 143–150 (2001)
24. Livingston, M., Decker, J.: Evaluation of trend localization with multi-variate visualizations. IEEE Trans. Visualization & Computer Graphics **17**(12), 2053–2062 (2011)
25. MacEachren, A.M., Roth, R.E., O'Brien, J., Li, B., Swingley, D., Gahegan, M.: Visual semiotics & uncertainty visualization: An empirical study. IEEE Transactions on Visualization and Computer Graphics **18**(12), 2496–2505 (2012)
26. Mittelstädt, S., Keim, D.A.: Efficient contrast effect compensation with personalized perception models. Computer Graphics Forum **34**(3), 211–220 (2015)
27. Pandey, A.V., Krause, J., Felix, C., Boy, J., Bertini, E.: Towards understanding human similarity perception in the analysis of large sets of scatter plots. In: Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 3659–3669 (2016)
28. Ryan, G., Mosca, A., Chang, R., Wu, E.: At a glance: Pixel approximate entropy as a measure of line chart complexity. IEEE Trans. Visualization & Computer Graphics **25**(1), 872–881 (2019)
29. Schloss, K.B., Gramazio, C.C., Silverman, A.T., Parker, M.L., Wang, A.S.: Mapping color to meaning in colormap data visualizations. IEEE Trans. Visualization & Computer Graphics **25**(1), 810–819 (2019)
30. Song, H., Szafir, D.A.: Where's my data? evaluating visualizations with missing data. IEEE Trans. Visualization & Computer Graphics **25**(1), 914–924 (2019)
31. Srinivasan, A., Brehmer, M., Lee, B., Drucker, S.M.: What's the difference?: Evaluating variations of multi-series bar charts for visual comparison tasks. In: Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 304:1–304:12 (2018)
32. Szafir, D.A.: Modeling color difference for visualization design. IEEE Trans. Visualization & Computer Graphics **24**(1), 392–401 (2018)
33. Tanahashi, Y., Leaf, N., Ma, K.L.: A study on designing effective introductory materials for information visualization. Computer Graphics Forum **35**(7), 117–126 (2016)

34. Volante, M., Babu, S.V., Chaturvedi, H., Newsome, N., Ebrahimi, E., Roy, T., Daily, S.B., Fasolino, T.: Effects of virtual human appearance fidelity on emotion contagion in affective inter-personal simulations. IEEE Trans. Visualization & Computer Graphics **22**(4), 1326–1335 (2016)
35. Yang, Y., Dwyer, T., Goodwin, S., Marriott, K.: Many-to-many geographically-embedded flow visualisation: An evaluation. IEEE Trans. Visualization & Computer Graphics **23**(1), 411–420 (2017)
36. Zhao, H., Bryant, G.W., Griffin, W., Terrill, J.E., Chen, J.: Validation of splitvectors encoding for quantitative visualization of large-magnitude-range vector fields. IEEE Trans. Visualization & Computer Graphics **23**(6), 1691–1705 (2017)
37. Zheng, L., Wu, Y., Ma, K.L.: Perceptually-based depth-ordering enhancement for direct volume rendering. IEEE Trans. Visualization & Computer Graphics **19**(3), 446–459 (2013)