Netspar

Network for Studies on Pensions, Aging and Retirement

NETSPAR ACADEMIC SERIES

The Comprehension and Processing of Text and Image Relationships

The influences of text-image relationships on text-image combined presentations

Hella de Jong

MSc 01/2020-015

The Comprehension and Processing of Text and Image Relationships

The influences of text-image relationships on text-image combined presentations.

Hella de Jong

SNR 2029742

Master Thesis

Communication and Information Sciences

Specialization Communication and Cognition

School of Humanities and Digital Sciences

Tilburg University, Tilburg

Supervisor: Dr. R. Cozijn

Second Reader: Dr. N. Cohn

January 2020

Abstract

Text and images are all around us and are used in various ways, for instance for movie posters or advertising campaigns. How text and images are semantically related to each other has been studied extensively, as well as how people process and comprehend them. Another perspective is to look at how text and images structurally relate to each other. They can be grouped into four text-image relationships: inherent, emergent, adjoined, and independent. The present study investigated what the influence of text-image relationships on the processing and comprehension of text-image combined presentations was with the manipulation of the congruence of the text. Participants had to judge whether text-image combinations, with statements placed according to the relationship in or near the image, as true or false. This resulted in that congruent text-image relationships were processed faster and comprehended better than incongruent text-image relationships. Congruent text-image relationships did not differ from each other regarding processing, but incongruent text-image relationships did. This might be explained by the integration process of the text and the image, as this should be easier and faster when the text and the image convey the same meaning. Whether this case holds if the cultural background of the participants is taken into account would be a suggestion for future research. As cultural background influences, for instance, the way people perceive color and therefore this might also influence the comprehension and processing of text-image combined presentations.

Keywords: text, image, comprehension, processing, relationships, integration

Table of Contents

The Comprehension and Processing of Text and Image Relationships

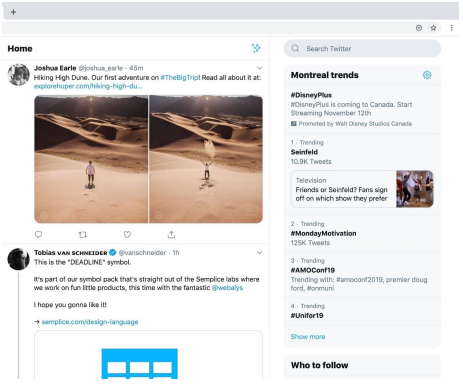
Text and images are ubiquitous in everyday life. Even back in prehistoric times, people used clay to draw representations of animals, hunting practices, and people on the inside of caves. Nowadays, text and images are present in almost any field, as Figure 1 illustrates through two examples. In A, Times Square displays advertisements of brands such as Coca Cola or more luxurious brands such as Tommy Hilfiger. In example B, posters and flyers are used, for example, to inform people about concerts, activities, or events in general.

A 

B 

Figure 1. Examples of text and image: Times Square street view (A) and street posters of events (B). Obtained from Unsplash.com (free images).

Furthermore, magazines portray colorful combinations of images and text, television uses text, images, and videos to entertain, inform, or educate people and text and images are used in infographics to inform people. For instance, within the latter, the number of participants might be displayed in the form of a pie chart instead of just regular text. Additionally, text and images are simultaneously used in the field of education. Lecture slides include visual elements to demonstrate certain topics or ideas, textbooks include images of theoretical models which are explained in the text and picture books for small children include a story, written in text, and are accompanied by visuals to enhance the story. Websites, for instance Tedx.com, use images, videos, and text to create an interactive webpage to entertain and educate people. The same goes for social media platforms, such as Instagram and Twitter, which are visually informative, but also include textual parts. Figure 2 displays the news feed of Twitter (A) and the Netflix interface (B), showing both text and images.

A 

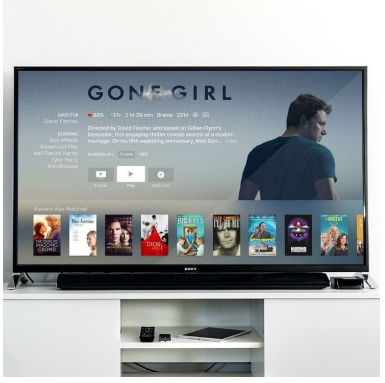
B 

Figure 2. Examples of text and image on the internet: Twitter news feed (A) and Netflix interface (B). Obtained from Unsplash.com (free images).

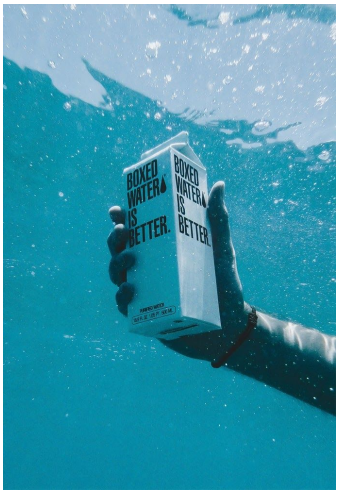
For instance, Netflix displays a still image of a movie together with the movie title and a short description with information about the movie. Similarly, mobile applications such as Thuisbezorgd.nl present images and textual parts of restaurants and their dishes and packaging of foods display the brand name, additional information, and very often an image.

Research has been done on what text and images are (e.g., Bateman, 2014), how they are in a relationship with one another (e.g., Schwarcz, 1982; Carney and Levin, 2002; Wu, 2014) and how text and images are processed and comprehended (e.g., Clark and Paivio, 1991; Rayner et al., 2001; Schnotz, 2002). These studies looked at various relationships between text and images as well as the integration of text and image, or the extent to which the text and image extend, complement, or enhance each other semantically. However, how text and image relate to each other in a structural way has not been investigated extensively.

According to Cohn (2013), there are four ways to integrate text and images structurally. These four structural text-image relationships are: inherent, emergent, adjoined, and independent (which are exemplified in Figure 3). An inherent text-image relationship is a text and image combination where the text and image are very closely related and part of each other's structure. The text is inherent in the object that is depicted in the image. The Boxed Water advertisement (3A) is an example of an inherent relationship. The advertisement displays a boxed water carton with the text "Boxed water is better" written on the carton. The second text-image relationship is emergent. Here, text and image are also closely related as well and some kind of visual affix is used to indicate this relationship. The

street art picture (3B) shows a girl with two text balloons saying “I felt the pressure building up! Pushing from every side! And you didn’t even mind”. The text balloon connects the textual part to the visual part. The text or speech from the balloon emerges from the object in the image, in this case, the girl in the picture. The third, an adjoined text-image relationship is similar to an emergent relationship but does not use visual affixes. The Heinz advertisement (3C) includes a ketchup bottle along with the slogan “No one grows Ketchup like Heinz”. The idea here is that the text and image are juxtaposed. Lastly, in the independent text-image relationship the text and the image seem unrelated, but are linked to each other by reference. For instance, (3D) shows a book page with a text and an image. Within the text, there is a reference between brackets “Figure 9.4” which refers to the image later in the text.

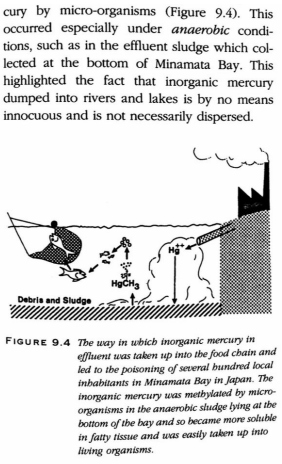
A



B

C

D



cury by micro-organisms (Figure 9.4). This occurred especially under *anaerobic* conditions, such as in the effluent sludge which collected at the bottom of Minamata Bay. This highlighted the fact that inorganic mercury dumped into rivers and lakes is by no means innocuous and is not necessarily dispersed.

FIGURE 9.4 *The way in which inorganic mercury in effluent was taken up into the food chain and led to the poisoning of several hundred local inhabitants in Minamata Bay in Japan. The inorganic mercury was methylated by micro-organisms in the anaerobic sludge lying at the bottom of the bay and so became more soluble in fatty tissue and was easily taken up into living organisms.*

Figure 3. Examples of different types of combining text and image: (A: Inherent) Boxed Water (Obtained from Unsplash.com, free images), (B: Emergent) Street art (Obtained from Unsplash.com, free images), (C: Adjoined) Heinz (Obtained from awards.brandingforum.org/brans/heinz/) and (D: Independent) a page out of a textbook (Timbrell, J. (2009). Introduction to Toxicology. p. 139).

In short, text and images are displayed all around us and are used for multiple purposes as mentioned before. However, the structural relationship between text and image has not been studied extensively. Cohn's (2013) theory suggests that there are four different text-image relationships and that they are comprehended and processed differently. But, this has not been tested in experimental research yet. The present study intends to fill this gap and seeks an answer to the question of what the influence of text-image relationships is on the processing and comprehension of text-image combined presentations.

Theoretical Framework

This chapter will elaborate further on the previously mentioned concepts. Text and images will be elaborated on as well as their relationships. Also, some light will be shed on the comprehension, integration, and understanding of text and images, as well as their relationships.

Text and image

Text can consist of individual letters, words, sentences, paragraphs, or even whole books. Someone can read a piece of text, but can also listen to a text which is read aloud to them. Bateman (2014) distinguishes between written text that a person can read and text that a person can hear. For this study, the focus will be on the written text. Bateman (2014, p.15) defines text as "... visually realized instances of language use".

Photographs, posters, and paintings are obvious examples of images. Bateman (2014) describes prior research of the family tree of images. Here five types of images are distinguished namely: graphic (e.g., family pictures), optical (e.g., the reflection of mirrors), perceptual (e.g., visual information of everyday life processed by the brain), mental (e.g., dreams or memories), and verbal (e.g., metaphors or descriptions). In this study, the focus will be on graphic images. Graphic images include (a) pictures of, for example, a dog, (b) statues of, for instance, a famous explorer and (c) design, for example, a graphic design for a wallpaper. To narrow this down, the focus will be on pictures (a). In particular still images of objects which "... we gain access by direct, physical visual perception" (Bateman, 2014, p.15).

Text and image relationships

A lot of research has been conducted on different text and image relationships. Schwarcz (1982) emphasized that there were nine categories for text-image relationships: congruency, elaboration, specification, amplification, extension, complementation, alternation, deviation, and counterpoint. They can be grouped into two general categories, namely congruency and deviation. With congruency, the text and image are in a well-balanced relationship. The image complements the text or vice versa. For instance, an

image could display a cat and the text would contain information about cats. The second category, deviation, puts the two opposite of each other. Here the image deviates from the text in some way, or the image and the text are unrelated. For example, the text is about food and the image shows a drink.

Furthermore, Nikolajeva and Scott (2001), based upon previous research, classified two major categories to which text-image relationships exist in picture books. The first category is: consonant, symmetrical, and complementary text-image relationships. A consonant relationship means that the text and image are in accordance with each other. When the text and the image are in a symmetrical relationship, the text and the image contain the same information. Complementary text-image relationships are characterized by the fact that the text and the image complement each other in some way. For instance, one page of a picture book could display the picture of a dog and the other page of the book would have a story about that dog. The second text-image relationship category is: enhancement and counterpoint. When an enhancement is present, the text is intensified by the image or vice versa. For instance, one page could display a certain dog breed and the other page detailed information about that dog breed. Here the image is enhanced by the detailed text. If the text and the image have a counterpoint relationship, they do not have something in common and give different information (e.g., one page with an image of a dog and the other page provides information about cats).

Wu (2014) also elaborates further on the text and image relationships found in picture books. There are three basic text-image relationships: elaboration, extension, and enhancement. First, elaboration is a relationship ".. of similarity across semiotic modes while no new ideational element is introduced by the text or image." (Wu, 2014, p.1417). Elaboration is divided into three sub-types: (1) exposition, the image and text strengthen each other by repeating or reformulating their meaning (e.g., one page contains the picture of a dog and the other page information about that dog), (2) exemplification, the text exemplifies the image and vice versa, but in different levels of generality (e.g., one page contains text about a mother and the other page displays the whole family including the mother), and (3) homospatiality, text and image merge together to form one entity in the picturebook (e.g., the text of a story is about a drive on a long road, the end of the text "long road" will eventually merge with the visual display of the road). Second, an extension is a relationship where the text and image extend each other's meaning and represent it. Extension is divided into three

sub-types: (1) augmentation, the image extends or adds new meaning to the text or the other way around (e.g., one page displays an image about an apple and the other page gives information about apples), (2) distribution, this "... refers to the construction of activity sequences with juxtaposed images and text jointly." (Wu, 2014, p.1417) (e.g., when a bigger illustration covers the meaning of both the text and image), and (3) divergence, the meaning of either the text or the image contradicts (e.g., the image displays a cat and the text gives information about dogs). Third, enhancement, the meaning of either the text or the image is expanded through the other either spatially, temporally, or causally. For instance, the text of a story might explain that the character was dressed nicely and then there would be several close up images of the clothes of that character showing how the character was dressed.

Another study by Carney and Levin (2002) explains that images improve students learning when reading a text. They distinguish five functions of pictures in text that is: decorational, representational, organizational, interpretational, and transformational. First, decorational images are purely decorative and have no relationship to the text. Second, when images are representational they contain the same content or at least a part of the same content as the text. For instance, a text elaborates on different dog breeds and the image displays some of those dog breeds. Third, organizational images contain content that extends the textual part, such as the display of a hiking trail in a book on hiking trails. Fourth, images are interpretational when the images provide extra information to clarify the text. For instance, the phases of a logistic process can be visualized in the form of a timeline. Lastly, transformational images are images mnemonic devices to improve the learning of the reader. According to Carney and Levin (2002), only the decorational images do not enhance the learning of students. The other functions are beneficial to learning from text, ascending from most beneficial to least: representational, organizational, interpretational, and transformational.

Even though the above-mentioned studies are just a glimpse of the extensively studied relationships between text and images there is also overlap in the categories. A text-image relationship that three agree upon is the relation in which the text and image contain the same object or convey the same meaning: congruency (Schwarcz, 1982), consonant (Nikolajeva and Scott, 2001), and exposition (Wu, 2014). In addition, when one looks at the bigger picture that Schwarcz (1982) has drawn, there is overlap with Nikolajeva and Scott (2001) regarding the categories complementation/complementary and counterpoint. Furthermore,

enhancement relationships occurred both in the studies of Nikolajeva and Scott (2001) and Wu (2014). As for the study by Carney and Levin (2002), representational images fall in the same category as congruency (Schwarcz, 1982), symmetrical (Nikolajeva and Scott, 2001), and exposition (Wu, 2014). In addition, organizational and interpretational images share features with complementary and enhancement (Nikolajeva and Scott, 2001), and extension and enhancement (Wu, 2014). Almost certainly, there will be more overlapping categories when it concerns text-image relationships in picture books as well as other fields where text and pictures occur side by side.

So, text and images are linked to each other in various relationships. However, these studies all looked at the semantic relationships between the text and images. A different approach is taken by Cohn (2013) who looks at the structural relationships between the text and the image. As explained earlier there are four ways text and image are mapped on to each other specifically: inherent, emergent, adjoined, and independent. An inherent relationship is a relationship between text and image where they are mapped into each other's structure. With emergent relationships, text and image are closely related by the use of visual affixes. Think of a panel in any comic strip where a character says something and the text is displayed in a text balloon. The third relationship is adjoined. Here the text and image are closely related but do not use any visual affixes. Imagine a poster of a product advertisement with the corresponding slogan of the product below the product. Lastly, the independent relationship is a relation where text and image seem totally separate. Often in textbooks images are included. The author might refer to an image in the text with the notion “see Figure X”. The image is then later displayed on the page. Visual examples of text-image relationships are given in Figure 3. Only, how are these different text and image relationships processed and understood by people?

Text and image integration, processing, and comprehension

Extensive research has been done on the comprehension of text and images and foremostly focused on the mnemonic functions of pictures in texts. Carney and Levin (2002) suggest that pictures in texts also improve the learning of students while reading a text. This was explained by Paivio’s dual coding theory (Clark and Paivio, 1991), which explains the psychological phenomena when someone processes and/or tries to comprehend images and

text. Clark and Paivio (1991) explain that visual and verbal components are both used to process and comprehend information. Text and images are processed differently on a cognitive level, which results in separate mental representations of the same information. These representations of information might be used later when someone recalls that information. For instance, someone might have stored the stimulus concept of a cat. Here the word cat, as well as the image of a cat, is stored in memory. If the person is asked to recall the information about the concept cat he or she will recall either the word cat, the mental image of the cat, or both simultaneously. When only the word cat is recalled, the mental image of the cat is not gone, it will probably come to mind later on or can be recalled at a different time. By having two different paths to recall information, the chances that someone remembers that information is increased which improves learning.

Moreover, images complement the text in a way that the text becomes more concentrated, compacit/concise, concrete, coherent, comprehensible, correspondent, and codable when pictures are added (Carney and Levin, 2002). Thus images might, for instance, complement the text and enhance it in such a way that it is easier to process, but where do people look at first, the image or the text? In the study by Rayner et al. (2001), half of the participants were asked to view a car advertisement and the other half a skin-care advertisement. The participants had to imagine they either needed to buy a car or skin-care products and had to obtain as much information as possible. Eventually, participants had to recall which advertisements they remembered and which ones they (dis)liked. It resulted in that participants spent more time viewing the text than at the picture part of the advertisement, regardless of the condition. They spent more time viewing the advertisement they were instructed to gain more information about. Hughes et al. (2003) confirmed this in their study in which students had to complete search tasks that involved text, images, and videos. They investigated if students would focus more on the text or on the images. Results showed that the textual parts were looked at longer than the picture. Surprisingly, a later study by Rayner et al. (2008) had contradicting results. In this eye-tracking study, participants viewed printed advertisements. Half of the participants were asked to rate the advertisements on likeability and the other half on effectiveness. Results showed that participants (in either the likeability or effectiveness condition) spent more time viewing the picture part of the advertisement than the textual part. Whether people look at images or text first it is also important to know how text-image relationships are processed and comprehended.

According to Cohn (2013), the comprehension and understanding of the text-image relationships differ in the way how they are structurally integrated with each other. According to this view, ascending from best understood and comprehended through lowest is the inherent text-image relationship first, followed by emergent, adjoined, and independent. The main idea here is that, for instance, with an inherent relationship the text and image and already part of each other's structure. Therefore, the person looking at the text-image combination does not need to cognitively infer they carry the same meaning. As proposed by the dual coding theory by Paivio (Clark and Paivio, 1991), the image of a cat and the word cat are already merged as one concept for the person looking. Therefore, the comprehension should be better and processing should be faster. Moreover, emergent relationships have the visual affix which indicates that the word and the image belong to each other and thus convey the same meaning. The adjoined relationship does not have that visual affix and thus is expected to be processed and comprehended less fast. With the independent relationship, the text and the image are separate, though they are related to each other by some sort of reference. For that reason, the textual and visual part are cognitively processed separately and will later form the same concept. In the view of the fact that the text-image relationships are processed and comprehended differently the research question for this study is stated as follows:

> *RQ:* What is the influence of text-image relationships on the processing and comprehension of text-image combined presentations?

Elaborating on this research question hypotheses are formulated, as the type of relationships will influence comprehension and processing (Cohn, 2013). In order to test this, the congruence of the text is manipulated. For instance, take the example of the image of a fish again. If the text "You can eat this" is placed next to the fish, placement of the text depends on the text-image relationship, this counts as congruent as one can eat fish. If the text "You can drive with this" is placed with to the fish, this counts as incongruent as one cannot drive a fish. As it matters where the text is placed relative to the image, the stronger text-image relationships should lead to better comprehension and faster processing. This would be in the case that the text and the image convey the same meaning, thus are congruent with each other. If the text and the image are incongruent with each other the order of better

comprehension and faster processing would be reversed, starting with independent, followed by adjoined, emergent, and inherent. However, the current study will follow the congruent text-image relationships and so the inherent text-image relationships should be processed the fastest and comprehended the best, followed by emergent, adjoined, and independent relationships. This leads to the following hypotheses:

H1: The stronger the text-image relationship, the faster it is processed and the better it is comprehended.

H1a: Inherent text-image relationships are processed faster and comprehended better than emergent, adjoined, and independent text-image relationships.

H1b: Emergent text-image relationships are processed faster and comprehended better than adjoined and independent text-image relationships.

H1c: Adjoined text-image relationships are processed faster and comprehended better than independent text-image relationships.

As mentioned before, in order to test which of the text-image relationships is processed the fastest and comprehended the best, congruence is manipulated. It is expected that text and images which are congruent are comprehended better and processed faster. For instance, if one sees the image of a fish with the text “fish”, the text and image convey the same meaning. However, if one sees the image of a fish with the text “skateboard” this causes possible confusion with the processing and comprehension. As the image and the text do not convey the same meaning. The manipulation of the congruency of the text-image relationships leads to two hypotheses:

H2a: Congruent text-image relationships are processed faster than incongruent text-image relationships.

H2b: Congruent text-image relationships are are comprehended better than incongruent text-image relationships.

Method

This chapter will elaborate on the experiment conducted for the current study. This includes: participants, materials, design, instrumentation, procedure, and data analysis.

Participants

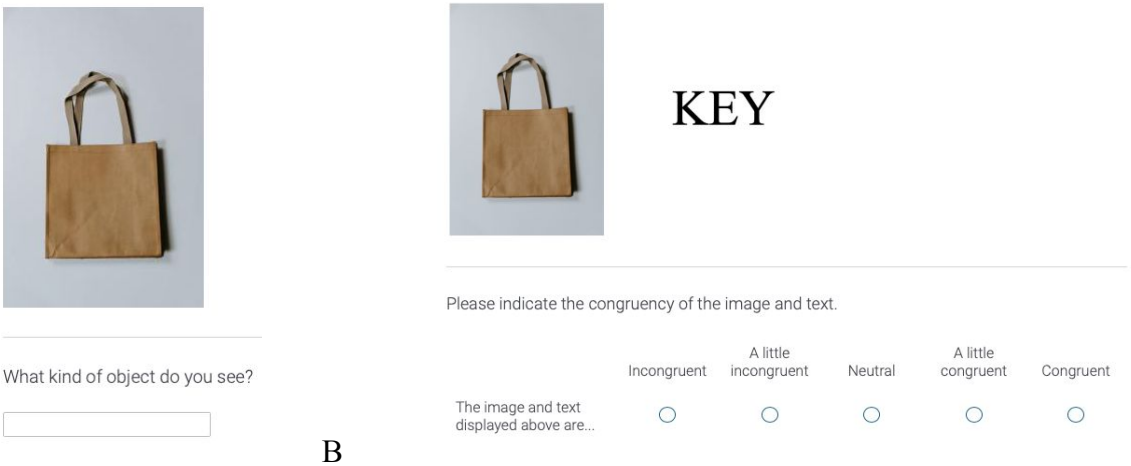
30 males and 55 females (*N* = 85) took part in the current study, with an average age of 23 years (*SD* = 3.46). Participants were approached by the researcher on the campus ground of Tilburg University and recruited via the Tilburg University Human Subject Pool. Participants from the Human Subject Pool who completed the survey were rewarded with 0.5 participant pool credit. 40 participants' highest level of education was a university of applied sciences, followed by pre-university education (21), university bachelor (13), university master (8), and secondary education (3). It was expected that the gender, age, country of residence, and education level of the participants would not influence the results. Therefore, no restrictions were set for the participants beforehand as the current study zoomed in on the processing and comprehension in general.

Materials

Pretest. To confirm that the objects in the images used in this study were viewed as the same by multiple people and the congruence of text and image was correct, a pretest has been performed. 30 participants, from the researcher's network, were approached to participate in a small survey with two main topic types: the validation of the images and the confirmation of the congruence.

29 images were obtained from unsplash.com a website that offers free images. 1 image, the "scarf" image, was made by the researcher as no suitable picture could be found on the website. The criteria for an image were as follows: (1) the object in the image should not be able to talk or make a sound (e.g., a telephone or television could not be included), (2) the name of the object in the image should consist of one syllable (e.g., car or chair), and (3) the object in the image should be easy recognizable. For the last requirement, the pretest included the picture validation part.

Participants for the pretest judged 30 images in total. Per image, the open question "What kind of object do you see?" was asked (see Figure 4A for an example). First, participants had to name what kind of object was displayed in the image. Second, participants were asked to indicate the congruence of an image and text combination. The list of 30 images was cut in half so that 15 images and 15 descriptions of the objects displayed in the image were used. The images and descriptions were randomly assigned to each other. So, participants viewed 15 text-image combinations, with on the left side the image and on the right a description of an object (see Figure 4B for an example). Participants were asked to indicate the congruency of the image and the text displayed on a 5-point Likert scale: (1) Incongruent, (2) A little incongruent, (3) Neutral, (4) A little congruent, (5) Congruent. Participants rated the 15 text-image combinations.



A B

Figure 4. Example of the picture validation question (A) and the congruence confirmation question (B).

As the picture validation questions were open questions coding criteria were set up: (1) articles were not coded (e.g., "the bag" was coded as "bag"), (2) different names were coded individually (e.g., "wine glass" and "glass of wine" were coded as two different names), and (3) if one did not answer in English, the word(s) were translated with the use of Google Translate and then coded. The results of the picture validation showed that the images: clock, book, tree, egg, watch, rose, and key had a 100% agreement rate among the participants. The images: lamp, door, glass, scarf, socks, and sun had a 96.7% agreement and leaf, pear knife, car, and ring had 93% agreement. The agreement percentage for the other images were: chair (90%), pen (90%), bag (86.7%), bread (86.7%), bench (83.3%), shoe

(73.3%), couch (70%), bed (66.7%), mug (66.7%), beer (66.7%), hand (56.7%), and wine (46.7%).

As for the congruence questions in the pretest, combinations were preferable rated as incongruent. Incongruent was rated as "1" on the 5-point Likert scale as described before. In Table 1, the means and standard deviations of the congruence scores of the text and images are displayed. The only exceptional case was bread//knife (*M* = 3.40, *SD* = 1.45). This could be caused by the fact that the bread and knife were strongly associated. Therefore, the bread/knife was excluded from the study.

Table 1

Means and standard deviations of the congruence scores on the text and images in the pretest. (1 = incongruent, 2 = a little incongruent, 3 = neutral, 4 = a little congruent, 5 = congruent)

Text and images	*M*	*SD*
leaf/socks	1.00	0
book/scarf	1.03	0.18
lamp/beer	1.03	0.18
pen/wine	1.03	0.18
bag/key	1.07	0.25
glass/egg	1.07	0.25
couch/shoe	1.10	0.31
chair/hand	1.13	0.57
mug/watch	1.13	0.57
tree/car	1.13	0.43
clock/sun	1.17	0.59
pear/bench	1.17	0.59
bed/rose	1.20	0.55
door/ring	1.20	0.66
bread/knife	3.40	1.45

Total	1.26	0.80

And so, the pretest resulted in 16 text-image combinations with an average picture validation score of 87.78% and an average congruence score of $M = 1.26$ ($SD = 0.80$).

Text and Image Comprehension. The materials used for the current study consisted of 128 text-image combinations. There were 16 stimuli sets in total. One set of stimuli consisted of the four text-image relationships (inherent, emergent, adjoined, and independent) with the two levels of congruence (congruent and incongruent). Below, Figure 5 displays an example stimuli set of the text-image relations in the congruent and incongruent conditions. The 16 stimuli sets used for the experiment are listed in Appendix A, categorized per item.

To conduct the experiment, several steps were taken. First, the 16 images that were used for the study were based on the agreement percentage of the pretest. The 16 images with the highest percentage were selected. This resulted in the images: bag, bed, book, car, clock, egg, glass, key, knife, lamp, leaf, rose, scarf, socks, sun, and tree. Second, the affordances of the 16 images were composed (see Appendix A). Affordances are characteristics of objects which distinguish that specific object from other objects which are similar (Gibson, 1977). For instance, an affordance of a bottle was that a person could drink from it. The most suitable affordance per image was chosen, based on logic and common sense as determined by the researcher and the supervisor. The affordance was converted into a suitable statement for the congruent condition. For the bottle, this resulted in the statement “You can drink from this”. The incongruent statement came from the counterpart image, which was tested in the pretest. As tested in the pretest, the counterpart image was incongruent, which resulted in an affordance, and so a statement, that did not match the image/text. The incongruent statement for the bottle example was “You can call this”, as someone cannot call a bottle. This process was repeated until the 16 images had a congruent and incongruent statement. Lastly, the 128 conditions needed to be created. With the use of canva.com, a free website where one can create and edit images, the conditions were created. Several criteria were used to control the creation of the text-image combinations: (1) the images kept their original width and height as these were tested in the pretest, (2) the images were placed in a 1000x1000 pixels white frame, (3) Open Sans Light was the font type and 16 was the font size, and (4) the text was in

bold and capitalized because this was more readable. Also, the default text color was black, but if the text was more readable in white the text color was changed to white. If the text color changed for one image condition, the other conditions for that stimuli set also changed text color. Moreover, for the inherent text-image relationship the text was placed on top of the object in the image. With the emergent images, the text was placed next to the object and an arrow was added to serve as a visual affix. In the adjoined relationships, the text was placed in the exact same place as with the emergent images, but with no visual affix. Lastly, in the independent condition, the text was placed below the image.

A1: Inherent congruent

A2: Inherent incongruent

B1: Emergent congruent

B2: Emergent incongruent

C1: Adjoined congruent

C2: Adjoined incongruent

D1: Independent congruent

D2: Independent incongruent

Figure 5. Example of congruent (1) and incongruent (2) stimuli in the four text-image relationships: (A1) Inherent congruent, (A2) Inherent incongruent, (B1) Emergent congruent, (B2) Emergent incongruent, (C1) Adjoined congruent, (C2) Adjoined incongruent, (D1) Independent congruent, and (D2) Independent incongruent.

Design

A 4x2 within-participants design was used to examine the hypotheses. The study consisted of two independent variables. The first was the text-image relationship with the

levels inherent, emergent, adjoined, and independent, and the second independent variable was the congruence with the levels congruent and incongruent.

The study consisted of eight lists of stimuli. Participants were randomly assigned to a list and stimulus were presented in a random order. A list consisted of the 16 images and were equally distributed over eight congruent and eight incongruent sets. For each congruence set, the images were again equally distributed over the four text-image relationships, which resulted in that each text-image relationship occurred twice. This meant that a participant would see 16 different images with twice the four text-image relationships in a congruent condition and twice the four text-image relationships in an incongruent condition. There were eight lists in total.

Instrumentation

The dependent variable for this study was the amount of time participants needed to judge the congruence of the text-image combination. In the section procedure, a detailed description will be given on how the participant had to judge the text-image combinations. Participants were instructed to view the text-image combination and indicate whether the text and image were congruent with each other. Instead of asking the participants “Are the text and image congruent?”, the participants were asked whether the statement provided was true or false (“Is this true or false?”). As this was a more natural way for participants to answer a question. Participants could answer the question with “True” by pressing the number 1 on their keyboard or with “False” by pressing the number 0 on their keyboard.

The time participants needed from the onset of the image on the screen up to the offset of the image on the screen was recorded in seconds. This time will be referred to as “reaction time”. The “lab.js” Javascript plugin for Qualtrics was used, which allowed for recording participant responses and the reaction times to those responses.

Furthermore, the accuracy of the participants was measured. So, whether the participants answered the questions correctly. For instance, the text-image combination of the bag/key was displayed. For the congruent condition of the bag, the statement “You can put something in this” was stated, here the correct answer was “True”. For the incongruent condition of the bag, the statement was “You can lock with this”, and so the correct answer was “False”. This same principle applied for the counterpart: the correct answer for the

statement in the congruent condition of the key “You can lock with this” was “True” and the correct answer for the statement in the incongruent condition of the key “You can put something in this” was “False”.

Besides the above measures, demographic information of the participants was asked and recorded. This included: gender, age, country of residence, and education level.

Procedure

Through the interface of the Tilburg University Human Subject Pool (ethical approval code: REC#2018/47, expiration date: 01/11/2020) participants received an invitation to a Qualtrics link which lead them to the survey. Or, if the participants were approached by the researcher, the Qualtrics survey was displayed on the screen of a laptop and placed before the participant on a table. The average time it took participants to complete the survey was four minutes.

First, the survey started with a welcome text and an informed consent form to which participants had to agree upon before starting the survey (see Appendix B). It was important for participants to complete the survey in one attempt as the experiment dealt with reaction times. However, the participants were not informed in detail that the study measured reaction times. The participants who were recruited via the Tilburg University Human Subject Pool, could read the following instructions in the preparation box on the website: “For this online study you will see some images with text and are asked to answer true/false statements. It is important to complete the survey on a laptop and in one go!! If not no credits will be assigned.”. If they had any further questions regarding the survey the email address of the researcher was provided on screen.

Second, participants were randomly assigned to a participant list and the actual experiment started. At this point, instructions were given how to complete the survey:

You will now see images and text statements.
You job is to answer the question “Is this true or false?”
If you think the statement is true, **press 1**.
If you think the statement is false, **press 0**.
Please press SPACE when you’re ready (you may need to click here with the mouse first!)

Participants would see one text-image combination at a time and were instructed if they thought the statement that went along with the image was true or false. Before every text-image combination, the participants were reminded for two seconds that the next text-image combination would appear on the screen soon, what the true/false keys were on the keyboard, and in which trail they were in, with the text:

Get ready for the next one!
TRUE = 1, FALSE = 0
This is trial 1 of 16

This repeated until the 16 text-images combinations from that participant list were judged. Third, the participants were asked to fill in their gender, age, country of residence, educational level and whether they spoke any other languages besides English. They were thanked, debriefed, and provided with contact information if they still had any questions regarding the experiment. Lastly, the participants had the opportunity to give their thoughts about the survey.

Data Analysis

As the experiment ran through an external server, coding was done to let the experiment run smoothly. In Appendix C the excel sheet used for the experiment is presented. First, the item number represents the object in the image (e.g., bag was “1”). Second, the scenario was indicated (e.g., bag). Third, the interface types were listed as well as the congruence. As there are four interface types (the four text-image relationships) and two levels of congruence, this resulted into eight possible options. Fourth, the text which was used for that particular image was listed (e.g., for the congruent bag image this was “you can put something in this”). Fifth, the corresponding image name was indicated (e.g., “bag_a_in_c” corresponded to the image of a bag with an inherent interface type and a congruent text). Sixth, to which participant list the image was assigned needed to be created. As a participant could only see one image of the stimuli set, eight lists were designed. From top to bottom, one to eight was counted and written down. When a new text-image

relationship was achieved, the list numbers were moved one place. This was repeated until the text-image combinations were listed. The lists now had 16 images, with every interface type twice (e.g., twice the inherent text-image relationship) and four times the congruent and four times the incongruent version. Lastly, to let the outside server read the document a last column was added to include the elements listed before into one cell at the end.

After the survey was conducted, the external server had collected the data and the data again had run through the javascript code, data preparation was done to run the processing (reaction time) analyses. First, an Excel sheet was created that included the eight conditions. The 85 participants were listed as rows, every participant response included the reaction times of the eight conditions and the accuracy of the eight conditions. Second, a score per condition per participant consisted of the average of two reaction times. If one item was answered incorrectly, the item was removed, which meant the score per condition per participant was based on one observation. 12.13% of the 1360 statements (85 participants * 8 conditions * 2 statements) were answered incorrectly. If there was no valid observation for a participant in a condition, the mean reaction time of that participant was substituted by the mean reaction time for the condition. Third, from the 680 data points (85 participants * 8 conditions) 16 average reaction times were considered outliers (2.4%). The cutoff point for outliers was calculated by 2.5 times the standard deviation on top of the mean. Five of the outliers were substituted with the average of that condition as both reaction times were above the before mentioned calculation and with 11 outliers the reaction time that was above the cut off which left the other observation as a replacement for the average.

Results

To investigate whether text-image relationships have an influence on the processing and comprehension of text-image combined presentations the data of 85 participants was analyzed. Chi-square analyses were conducted for the comprehension of the text-image relationships. Here 1360 data points (85 participants * 16 text-image combinations) were used. The percentages of correctly judged text-image combinations as a function of congruence and text-image relationships are listed in Table 2. In total, 95.0% of the congruent text-image combinations were answered correctly and 81.9% of the incongruent text-image combinations were answered correctly.

Table 2

Percentages of the correctly answered statements about the text-image combinations as a function of congruence and text-image relationships

	Congruent	Incongruent
Text-Image Relationship		
Inherent	98.2%	82.4%
Emergent	92.4%	78.8%
Adjoined	97.1%	84.1%
Independent	92.4%	82.4%

The test results of the Chi-square analysis revealed that congruent text-image combinations were statistically comprehended better than incongruent text-image combinations, χ^2 (1) = 50.452, $p < .001$. Therefore, hypothesis (H2b) congruent text-image relationships are comprehended better than incongruent text-image relationships, was supported. Analysis showed that the congruent text-image relationships were better comprehended than the incongruent text-image relationships: inherent, χ^2 (1) = 22.091, $p <$ .001., emergent, χ^2 (1) = 10.796, $p < .001$., adjoined, χ^2 (1) = 15.125, $p < .001$., and independent, χ^2 (1) = 6.721, $p = .010$.

Three one-way repeated measures ANOVA were conducted to analyze the processing (reaction times) of the text-image relationships. First, the congruence in general was

analyzed, second, the congruent text-image relationships, and third, the incongruent text-image relationships. The means and standard deviations of the reaction times of the text-image relationships in the congruent and incongruent conditions are presented in Table 3. On average for the congruent conditions, the independent text-image relationships had the fastest reaction time, followed by inherent, adjoined, and emergent. On average for the incongruent conditions, the independent text-image relationships again had the fastest reaction time, followed by adjoined, emergent, and inherent.

Table 3

Means and standard deviations of the reaction times of the text-image relationships separated by congruence.

	Congruent		Incongruent	
Text-Image Relationship	*M*	*SD*	*M*	*SD*
Inherent	2555	966	3725	2124
Emergent	2667	1372	3564	1892
Adjoined	2563	1010	3347	1344
Independent	2421	992	3075	1368
Total	2551	811	3428	1187

Note. Means and standard deviations are in milliseconds and rounded.

The first ANOVA for congruence type revealed a significant main effect of congruence type on the reaction times of the text-image relationships, $F(1,84) = 78.46$, $p < .001$, $\eta^2 = .483$. Text-image relationships in the congruent conditions were processed faster than text-image relationships in the incongruent conditions. Therefore hypothesis (H2a): congruent text-image relationships are processed faster than incongruent text-image relationships, was supported.

For the second ANOVA, Mauchly's test of sphericity indicated that the assumption of sphericity was violated for the main effect of congruent text-image relationships, χ^2 (5) = .810, $p = .004$. Therefore, the degrees of freedom were corrected with the use of the Huynh-Feldt estimates of sphericity ($\varepsilon = .92$). The results revealed that there was no

significant main effect of the text-image relationships in the congruent conditions, $F(2.759,231.727) = 1.185$, $p = .315$, $\eta^2 = .014$. These results were not in line with (H1) The stronger the text-image relationship, the faster it is processed and the better it is comprehended, for the processing part. In addition, repeated contrasts showed that inherent text-image relationships were not significantly processed faster compared to independent text-image relationships, $F(1,84) = 1.807$, $p = .182$, $\eta^2 = .021$. Adjoined text-image relationships were also not significantly processed faster compared to independent text-image relationships, $F(1,84) = 2.725$, $p = .103$, $\eta^2 = .031$. Emergent text-image relationships were also not significantly processed faster than independent text-image relationships, $F(1,84) = 1.326$, $p = .253$, $\eta^2 = .016$. Thus, hypotheses H1a, H1b, and H1c could not be supported for the faster processing part.

Also with the third ANOVA, Mauchly's test of sphericity indicated that the assumption of sphericity was violated for the main effect of incongruent text-image relationships, χ^2 (5) = .750, $p < .001$. Therefore, the degrees of freedom were corrected with the use of the Greenhouse-Geisser ($\varepsilon = .852$). The results showed that there was a significant main effect of the text-image relationships in the incongruent conditions, $F(2.555,252) = 3.297$, $p = .028$, $\eta^2 = .038$. Planned contrasts showed that independent text-image relationships were significantly processed faster than inherent text-image relationships, $F(1,84) = 8.743$, $p = .004$. Independent text-image relationships were also significantly processed faster than emergent text-image relationships, $F(1,84) = 5.466$, $p = .022$, $\eta^2 = .061$. Emergent text-image relationships were not significant processed faster than independent text-image relationships, $F(1,84) = 2.931$, $p = .091$, $\eta^2 = .034$.

In Figure 6, the reaction time data is shown of the inherent, emergent, adjoined, and independent text-image relationships in the congruent and incongruent conditions.

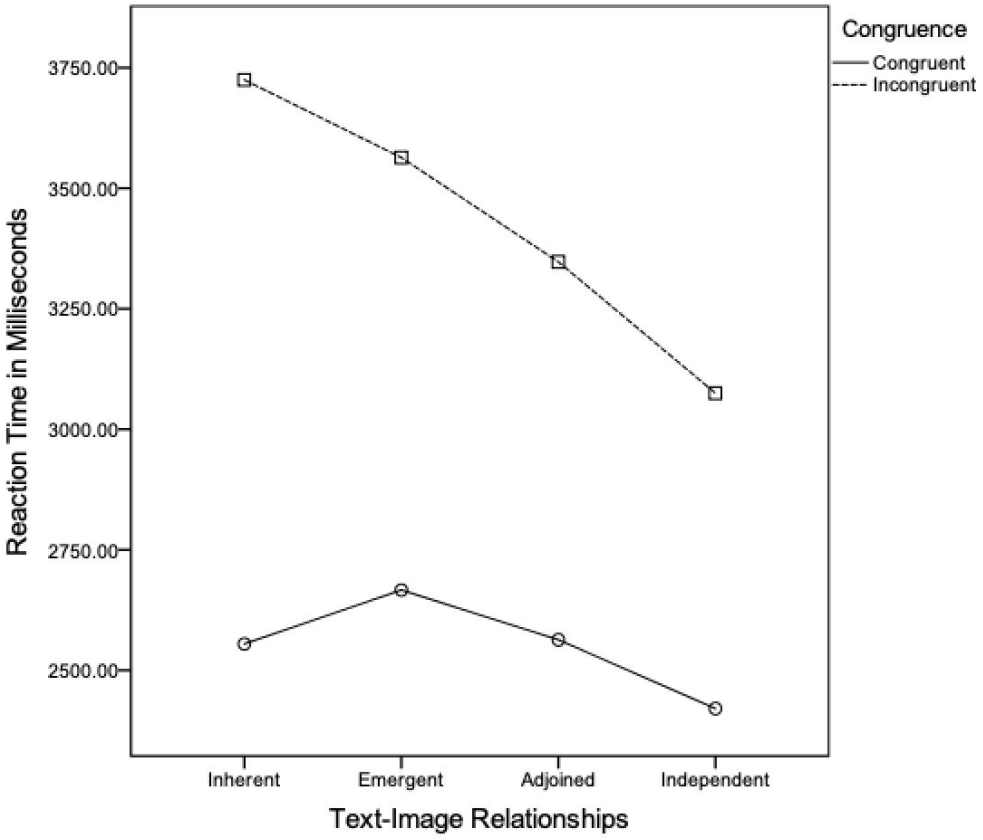


Figure 6. Reaction times of the text-image relationships as a function of relationship type and congruence.

Conclusion

The current study sought to investigate what the influence was of the text-image relationships on the processing and comprehension of text-image combined presentations. Studies extensively looked at the different semantic relationships between text and images. (e.g., Schwarcz, 1982, Nikolajeva and Scott, 2001, and Wu, 2014). The study by Cohn (2013) took a different approach and looked at the structural relationships between text and images. Cohn (2013) identifies four text-image relationships, namely inherent, emergent, adjoined, and independent. In the current study it was hypothesized that the stronger the text-image relationship, the faster it is processed and the better it is comprehended (H1). Analyses showed that when the text and the image were congruent with each other they did not statistically differ. Hypotheses H1, H1a, H1b, and H1c were not supported. However, analyses did show that when the text and the image were incongruent with each other, there was a statistical significant effect for congruence overall and for two of the three contrasts, indicating that the incongruent-independent relationships were processed faster than the incongruent-inherent and the incongruent-emergent relationships. Additionally, the results showed that the congruent text-image relationships were comprehended better and processed faster than the incongruent text-image relationships. Therefore, H2a and H2b were supported. Furthermore, in the congruent condition the inherent text-image relationship had the highest rate of correct answers, followed by adjoined, emergent, and independent. Although not statistically proven it does show that the inherent text-image relationships were on average comprehended better than the other text-image relationships.

The text-image relationships as proposed by the study of Cohn (2013) have an influence on the processing and comprehension of text-image combined presentations. When the text and the image are congruent with each other they are processed faster and comprehended better than when the text and the image are incongruent with each other. Thus one might say that when a statement about a visual object aligns with the meaning of the object, it is easier to judge their congruence than when the statement does not align. It was expected that the inherent text-image relationships would be processed the fastest, followed by the emergent, adjoined, and independent relationships. As can be seen from the reaction time averages of the text-image relationships and the data in Figure 6 this is not really the

case. The congruent independent text-image relationships had on average a faster reaction time compared to the other congruent text-image relationships. Similarly in the incongruent conditions, on average the independent text-image relationships were faster than the inherent text-image relationships and the emergent text-image relationships. Also, when the text and the image are congruent they are processed faster and comprehended better. Thus, the text-image relationships proposed by Cohn (2013) do have influence on the comprehension and processing of text-image combined presentations.

Discussion

The results of the current study might be related to Paivio's dual coding theory (Clark and Paivio, 1991). As concluded, the congruent text-image relationships were processed faster and comprehended better than the incongruent text-image relationships. To give an example, when participants saw the object in the image of a book they had to access their mental model of a "book", see if it matched the book in the image and conclude that the object in the image is indeed a book. Simultaneously, participants had to analyze the text that was placed next to the image of the book. In the congruent condition this would be "You can read this". Based on their previous knowledge on linguistics, semantics, and grammar, participants would analyze the sentence and give meaning to it, in this case, an affordance of a book, because one can read a book. Lastly, participants would integrate the two mental models into one, concluding that they convey the same meaning. At this point, the question "Is this true or false?" needed to be answered, which was "Yes" for the congruent conditions as the text and the image conveyed the same meaning. On the other hand, if the text stated "You can wear this" there would be a disruption. Since one cannot wear a book as one would wear a scarf, for example. Generally speaking, the affordance "to wear something" is not part of the meaning structure of a book. This means that when participants analyze the sentences in the incongruent condition and are the point of integration, the disruption causes confusion which leads to a longer reaction time and in some cases statements answered incorrectly, because participants had to conclude that the image and the text do not convey the same meaning and thus cannot be integrated. This would explain why a match between text and image facilitates their combined processing.

Other factors may also play a role in how fast the participants judged the text-image combination statements. Participants were instructed to to complete the experiment in one go, besides having to finish it on a laptop. However, these instructions for the completion did not include that the participants had to give their judgements as fast as possible. The question then is whether the reaction times sampled in the study reflect fast processing. One can expect that if it is explicitly stated that participants must respond as quickly as possible the average reaction times in the conditions will be somewhat faster. Besides this, faster processing could also be elicited by inclusion of a practice round. However, the current study did not include one. Practice might give the participants a feeling about is expected from

them. Participants maybe first needed time to adjust to the task before they would be able to focus on the experiment and judge the congruency statements.

Next to the congruence, this study looked at the influence of text-image relationships as proposed by Cohn (2013). The congruent text-image relationships were comprehended better and processed faster than the incongruent text-image relationships. When one looks at the average reaction times for congruent text-image relationships, the average of the independent text-image relationships is somewhat faster than the rest. As displayed in Figure 6, there is a little rise in the average reaction time for congruent-emergent text-image relationships. When comparing, the emergent and the adjoined text-image relationship in this study differed with only an arrow pointing from the text towards the object in the image in the emergent condition. It might be that when participants recall information, that there is a cognitive disruption when seeing the arrow going from the text to the object. Maybe for the emergent condition there were extra steps involved, like identifying the arrow, linking it to both the text and the image, analyzing which way it is pointing, concluding that the arrow is part of the whole image and indicates the relation between the text and the image. Only after all these steps, the viewer is able to judge the statement as true or false. Given this suggestion, the average of the incongruent adjoined text-image relationships is also slightly faster than the incongruent emergent text-image relationships.

The theory by Cohn (2013) suggest that the stronger the text-image relationships the better the comprehension and the faster processing. The hypotheses in this study were formulated for the congruent condition. From a congruent point of view, the expectation of better comprehension and faster processing would be inherent, emergent, adjoined, and independent, in descending order. If the hypotheses were viewed from an incongruent perspective, however, one might expect the direction to be reversed: independent, adjoined, emergent, and inherent. The results for the incongruent text-image relationships showed that there was a significant main effect. Independent text-image relationships were processed faster than inherent relationships as well as emergent relationships. To link this to Paivio's dual coding theory (Clark and Paivio, 1991). If the text and the image do not belong together, such as in the incongruent condition, it may be easier to come to the judgement to treat them separately than if they are presented together. The structurally overlapping representation may make it more difficult to conclude that the two parts do not belong together. When, for instance, the participants saw the image of knife they accessed their mental model a knife. At

the same time, analyze the piece of text which was placed next to the object, on the object, or below the image. The question “Is this true or false?” along the incongruent statement for this example “You can illuminate with this” should result in the answer “False”. Participants do not need to integrate the two mental models. Which could mean that it is easier to identify the two elements, the text and the image, separately than stronger structurally related. Which was the case in the inherent, emergent, and adjoined text-image relationships. Here, the text was structurally mapped on the object (inherent), next to the object with use of an arrow as a visual affix (emergent), and next to the object with no arrow (adjoined).

Given the structural mapping on the object and next to the object in the image presented, factors such as font size, font color, or a titled font might also have an influence. In the current study, the text of eight of the stimuli sets (50%) had a black font color and the other half had a white font color. Of the 16 stimuli sets, in 7 stimuli sets the text of the inherent images had to be titled in order to be properly mapped on to the object. In the case of inherent images, participants might have had difficulty reading the text that was displayed. This may have increased the reaction time for this condition. For the emergent and adjoined images, in 12 stimuli sets the text was placed in top right of the object, in 2 sets top left, in 1 set bottom right, and in 1 set middle right. Given that the participants Given that the participants all were accustomed to reading from left to right, this might have led them to view the text-image combinations from top left to bottom right, as they are used to when reading. A suggestion could be to fix the text at one position, whether it be top left/right, middle left/right, or bottom left/right. Moreover, with the help of eye-tracking technology, research may analyze where people look at first when presented with various text-image combinations. The font size was for all stimuli sets the same except for 2 stimuli sets. For the inherent images the font size needed to be adjusted in order to fit on to the object in the image. It could be the case that for those stimuli, the text was not readable and participants randomly hit 1 or 0 as they could not read the text properly. The data shows that for the image of the sun in the congruent condition no mistakes were made, but in the incongruent condition six mistakes were made. The other stimuli of the key only one mistake was made in the incongruent condition. Note should be given to that participants might guess correctly even though they could not read the text properly. In addition, the text was capitalized as this improved the reading of the text. However, capitalized letters are generally interpreted as an

increase of the tone of voice or to emphasize certain words in a sentence, for example. Depending on the background of the participant, it might even be considered offending.

A pretest was conducted to validate the objects in the images and to confirm the affordances of the objects. The image of the glass, for example, might be interpreted by one as a drinking glass, but by others as a tea or a coffee glass. The affordance of such a glass might also differ among people. As all the images, except one, were collected from a website which provides free images (unsplash.com), careful selection was needed as the website offers many images when searching for “glass”. Differences in the interpretation of the objects displayed in the images might also have had a share in the outcomes, since participants had to access their mental model the object in order to judge the statement that was presented. It could be that, because some participants might interpret the object slightly differently this process took longer because participants could not identify the object immediately. In addition, six sets of stimuli displayed additional visuals, such as a painting, a hand, or a road. As this might distract people when viewing an image, it could be preferred to have solely the object in the image with a plain background, and nothing else. Even the backgrounds differed among the images, the image of the egg had a plain red background while the image of the leaf had a blurred forest. All of these concerns may have had an influence on the recognizability of the object and thus on the speed with which the congruence could be judged.

Given the suggestions above, for future research it might be interesting to further analyze the influences of these structural text-image relationships as proposed by Cohn (2013). In the scope of this study, it can be concluded that there is an influence of these text-image relationships. It would take more time to identify the reasons to how people view text and images, how they identify the different structural relationships, and how text and the image are integrated given their structural relationship

Limitations and Future Research

The current study did not incorporated the cultural background of the participants as the focus was on the overall comprehension and processing of people. The objects in the images used in this study had different colors, shapes, and sizes, and according to Pettersson (1982) the perception of images and colors is affected by the cultural and geographical background of people. The study concluded that, for instance, illustrations in picture books differs among various cultures. For example, there were differences in pictorial illustrations, photographs, and whether these two were displayed in full color or in black and white. The participants in the current study studied at Tilburg University and spoke English, but could also speak other languages which might indicate different cultural backgrounds. Besides English, participants could speak Dutch (56), German (22), Spanish (9), French (5), Portuguese (3), Turkish (2), Papiamento (2), Bulgarian (2), Russian (2), Arabic (1), Chinese (1), Latin (1), Hungarian (1), Italian (1), Polish (1), Romanian (1), Japanese (1), Latvian (1), and the other 30 participants indicated either "No" or did not fill in the question whether they spoke another language. Pettersson (1982) stated that there are differences in color perception, color preferences, and color naming. For instance, in some languages (e.g., African) there is only a differentiation made between "light" and "dark" instead of specifically naming the color. Therefore, it would be interesting to see whether cultural background has an influence on perception of images and thus eventually the comprehension and processing of images.

Besides the cultural background of the participants, also age was not incorporated in the current study. As mentioned before, all participants studied at Tilburg University and therefore it can be assumed they all fall in the age category of +/- 18 years up to +/- 30 years. Generally speaking, older people have more knowledge than younger people, but does this also influence their comprehension and processing of information? The study by Zabrucky, Moore, and Schultz (1987) used an error-detection paradigm to analyze whether older adults differed from younger adults in the case of text comprehension. Their results showed that these two groups did not differ concerning their age. Only how information was offered in 1978 compared to now will probably differ, as the internet, for example, only existed for five years when the study was published. Nowadays, people are given multiple options to gain

information, textual and visual. For instance, via social media platforms, advertising posters, or (educational) videos. To link this to the text-image relationships by Cohn (2013) it might be of interest to see whether age mediates, moderates, or has an influence at all on the comprehension and processing of the text-image relationships. As older people have more life experience and thus probably also more knowledge than younger people. In addition, the way how information is presented to people in general nowadays might also influence the way people process and comprehend different text-image combinations.

References

Bateman, J. A. (2014) *Text and Image: A critical introduction to the visual/verbal divide* (1st edition). Abingdon, United States: Routledge.

Carney, R., & Levin, J. (2002). Pictorial Illustrations Still Improve Students' Learning From Text. *Educational Psychology Review, 14*(1), 5-26. doi: 10.1023/A:1013176309260

Clark, J. M., & Paivio, A. (1991) Dual Coding Theory and Education. *Educational Psychology Review*, 3(3), Plenum Publishing Corporation

Cohn, Neil. 2013. Beyond word balloons and thought bubbles: The integration of text and image. *Semiotica*. 2013(197), 35-63. doi: 10.1515/sem-2013-0079

Gibson, J.J. (1977) The theory of affordances. In R.E. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 56-60). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hughes A., Wilkens T., Wildemuth B.M., Marchionini G. (2003) *Text or Pictures? An Eyetracking Study of How People View Digital Video Surrogates.* CIVR 2003. vol 2728. doi: 10.1007/3-540-45113-7_27

Nikolajeva, M. & Scott, C. (2001). *How picturebooks work.* New York, NY Garland.

Pettersson, R. (1982). Cultural Differences in the Perception of Image and Color in Pictures. *Educational Communication and Technology, 30*(1), 41-53. Retrieved January 24, 2020, from www.jstor.org/stable/30219824

Rayner, K., Miller, B., Rotello, C. M. (2008). Eye Movements When Looking at Print Advertisements: The Goal of the Viewer Matters. *Applied Cognitive Psycholog*y, 22, 697-707. doi: 10.1002/acp.1389

Rayner, K., Rotello, C. M., Steward, A. J., Keir, J., Duffy, A. (2001). Integrating text and Pictorial Information: Eye Movements When Looking at Print Advertisements. *Journal of Experimental Psychology: Applied*, 7(3), 219-226. doi: 10.1037//1076-898X.7.3.219

Schnotz, W. (2002). Towards an integrated view of learning from text and visual displays. *Educational Psychology Review,* 14(1), 101-120. doi: 10.1023/A:1013136727916

Schwarcz, J. (1982). Ways of the illustrator: Visual communication in children's literature. Chicago, IL: American Library Association.

Wu, S. (2014). A Multimodal Analysis of Image-text Relations in Picture Books. *Theory and Practice in Language Studies*, 4(7). doi: 10.4304/tpls.4.7.1415-1420

Zabrucky, K., Moore, D., & Schultz, N. R. (1987). Evaluation of comprehension in young and old adults. *Developmental Psychology, 23*(1), 39–43. doi:10.1037/0012-1649.23.1.39

Appendices

Appendix A

Images and Texts

In this appendix, the 16 stimuli sets are presented which were used for this study, categorized by item. On the left side the original image including the white borders and on the right side of the image are the pretest results and the congruent and incongruent statements used for that stimuli set. Below are the eight conditions displayed, going from left to right: inherent, emergent, adjoined, and independent. For the upper row the text is congruent with the image, the lower row the text is incongruent with the image.

1. Bag

Pretest results

Picture validation: bag (86.67%), paper bag (13.33%), shopping bag (3.33%)

Congruence: $M = 1.07$, $SD = 0.25$

Statements

Congruent: You can put something in this.

Incongruent: You can lock with this.

Experimental conditions

2. Bed

Pretest results

Picture validation: bed (66.67%), bedroom (26.67%), art (3.33%), bed and canvases (3.33%)

Congruence: $M = 1.20$, $SD = 0.55$

Statements

Congruent: You can sleep in this.

Incongruent: You can pick this.

Experimental conditions

3. Book

Pretest results

Picture validation: book (100%)

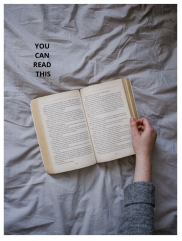
Congruence: $M = 1.03$, $SD = 0.18$

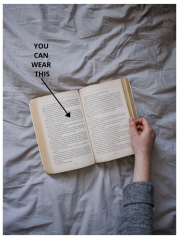
Statements

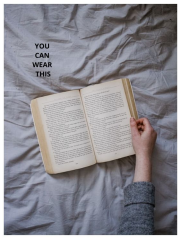
Congruent: You can read this.

Incongruent: You can wear this.

Experimental conditions







4. Car

Pretest results

Picture validation: car (93.33%), audi (6.67%)

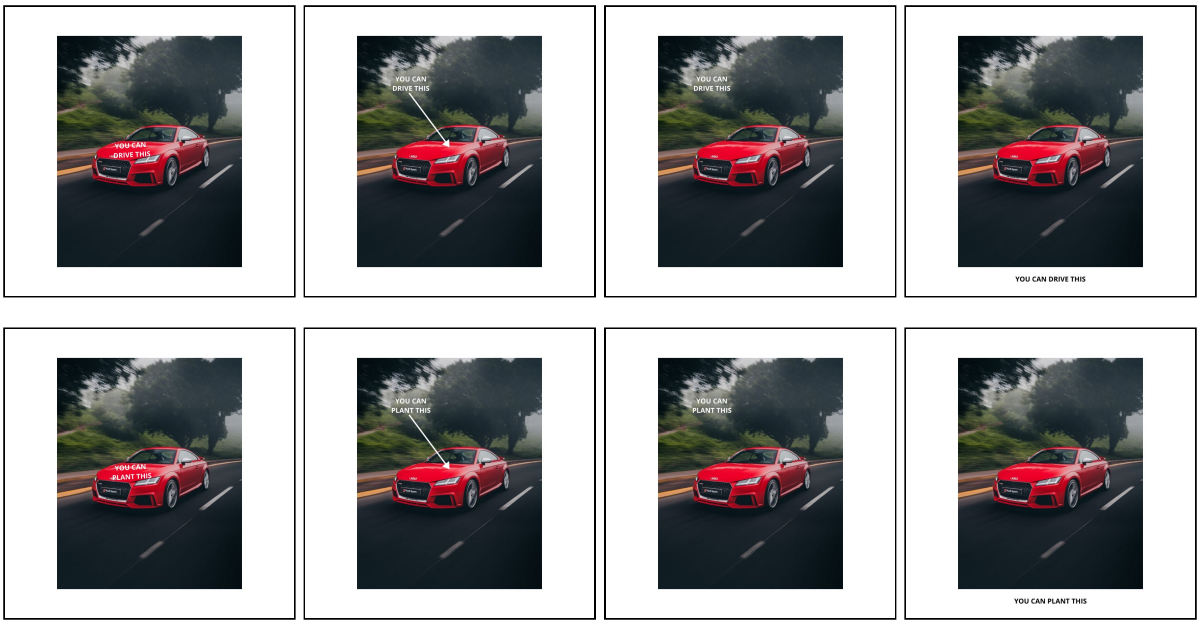
Congruence: $M = 1.13$, $SD = 0.43$

Statements

Congruent: You can drive this.

Incongruent: You can plant this.

Experimental conditions



5. Clock

Pretest results

Picture validation: clock (100%)

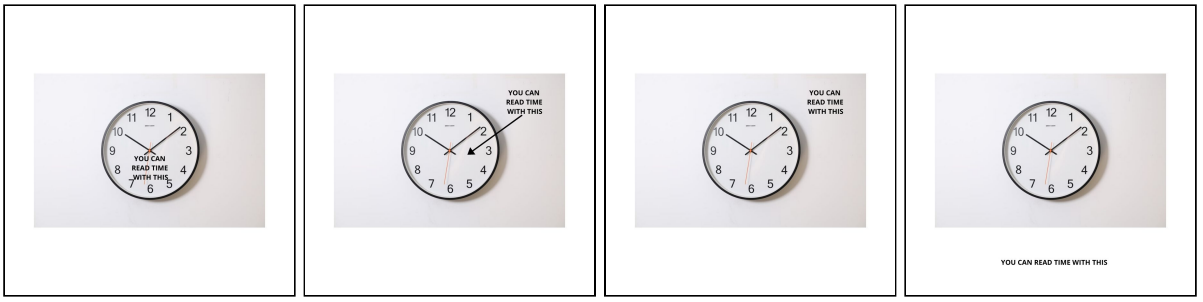
Congruence: $M = 1.17$, $SD = 0.59$

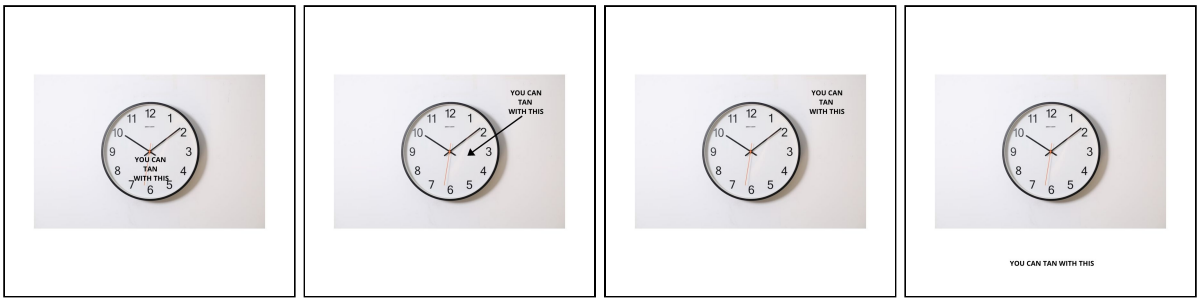
Statements

Congruent: You can read time with this.

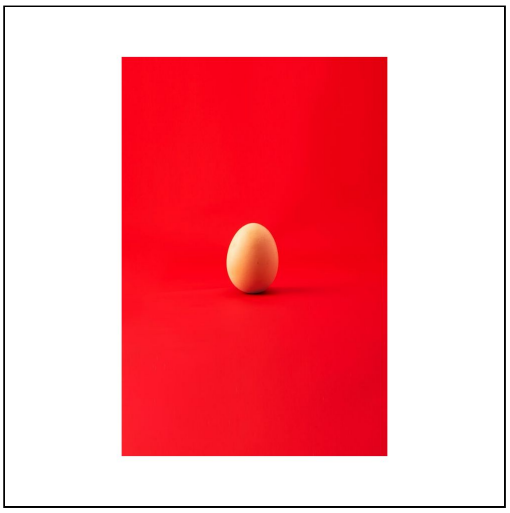
Incongruent: You can tan with this.

Experimental conditions





6. Egg



Pretest results

Picture validation: egg (100%)

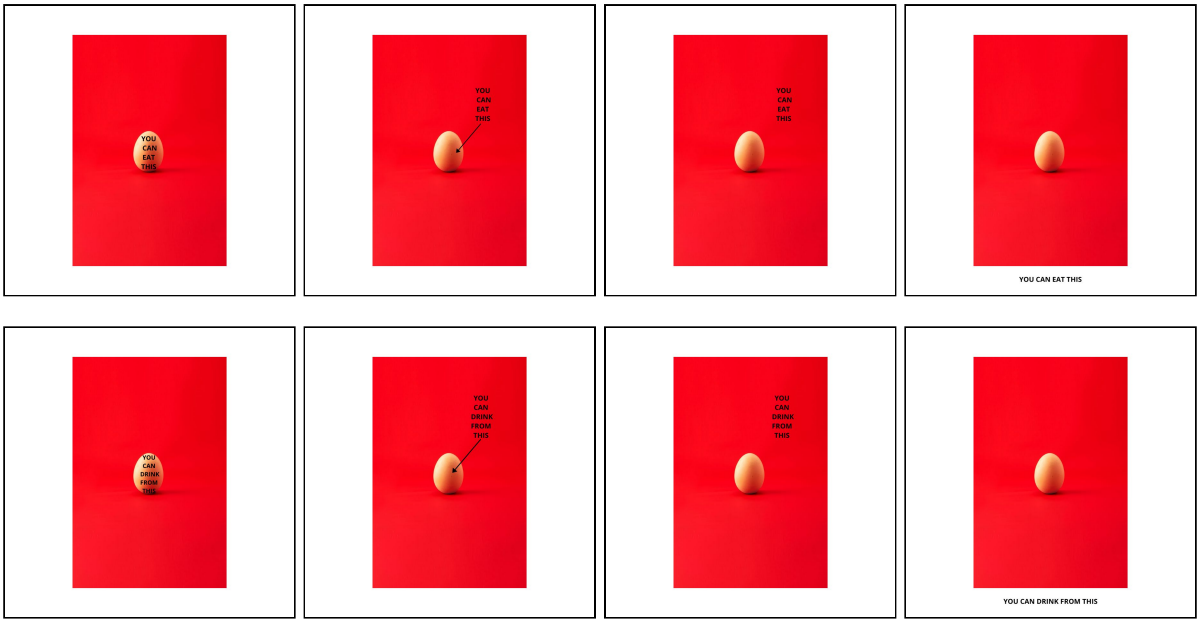
Congruence: $M = 1.07$, $SD = 0.25$

Statements

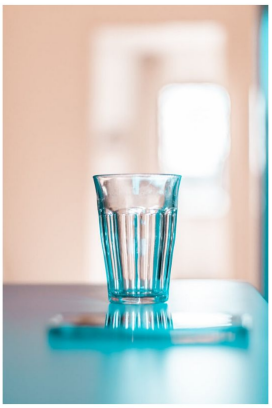
Congruent: You can eat this.

Incongruent: You can drink from this.

Experimental conditions



7. Glass



Pretest results

Picture validation: glass (96.67%), drinking glass (3.33%)

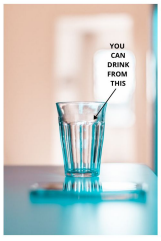
Congruence: $M = 1.07$, $SD = 0.25$

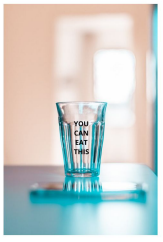
Statements

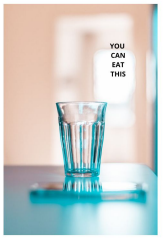
Congruent: You can drink from this.

Incongruent: You can eat this.

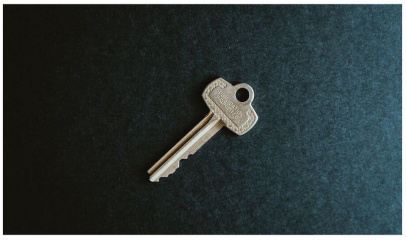
Experimental conditions







8. Key



Pretest results

Picture Validation: key (100%)

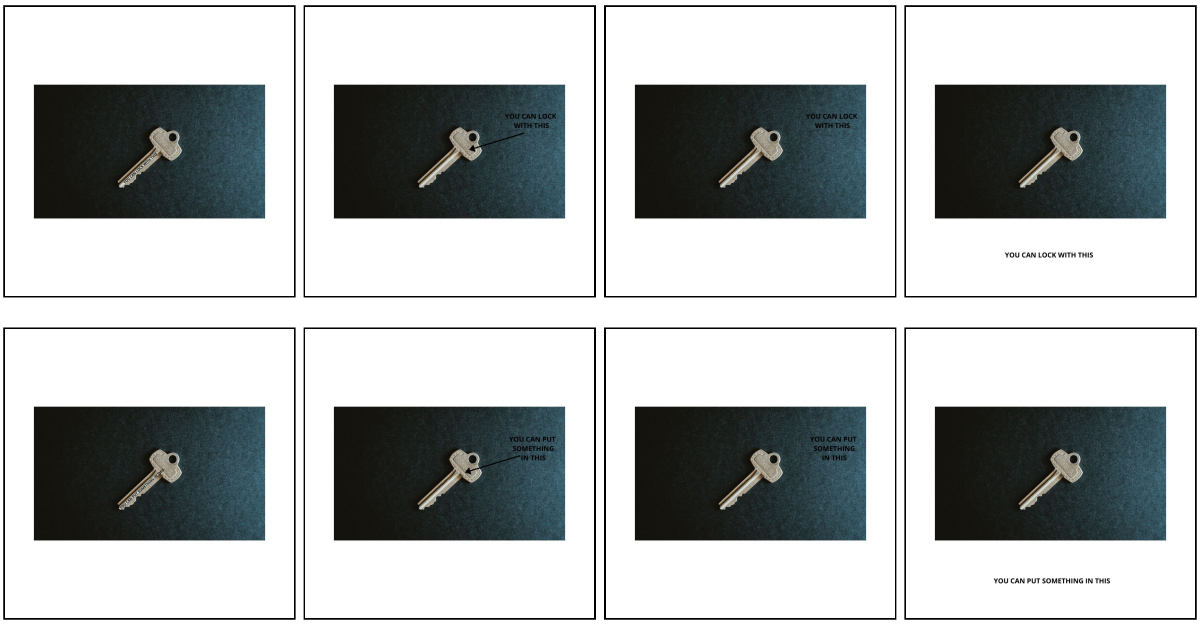
Congruence: $M = 1.07$, $SD = 0.25$

Statements

Congruent: You can lock with this.

Incongruent: You can put something in this.

Experimental conditions



9. Knife

Pretest results

Picture validation: knife (93.335), kitchen knife (3.33%), cutting knife (3.33%)

Congruence: n.a.

Statements

Congruent: You can illuminate with this.

Incongruent: You can cut with this.

Experimental conditions

10. Lamp

Pretest results

Picture validation: lamp (96.67%), standing lamp (3.33%)

Congruence: n.a.

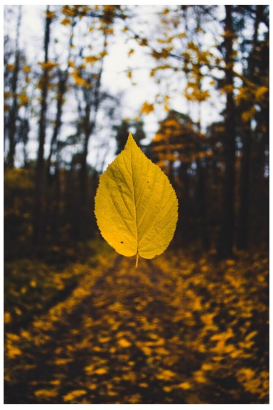
Statements

Congruent: You can illuminate with this.

Incongruent: You can cut with this.

Experimental conditions

11. Leaf



Pretest results

Picture validation: leaf (93.33%), autumn leaf (3.33%), leaf and trees (3.33%)

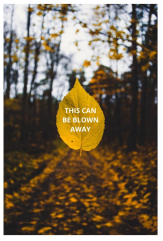
Congruence: *M* = 1.00, *SD* = 0.00

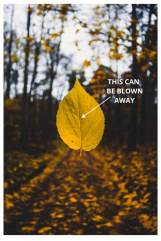
Statements

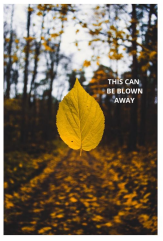
Congruent: This can be blown away.

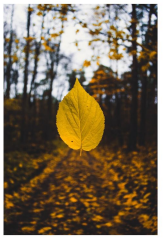
Incongruent: You can put these on.

Experimental conditions

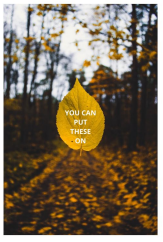


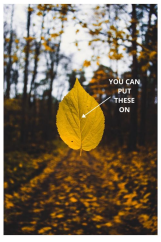


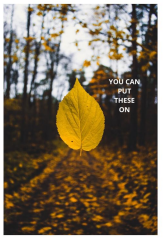


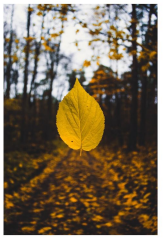


THIS CAN BE BLOWN AWAY









YOU CAN PUT THESE ON

12. Rose

Pretest results

Picture validation: rose (100%)

Congruence: *M* = 1.20, *SD* = 0.55

Statements

Congruent: You can pick this.

Incongruent: You can sleep in this

Experimental conditions

13. Scarf

Pretest results

Picture validation: scarf (96.67%), . (3.33%)

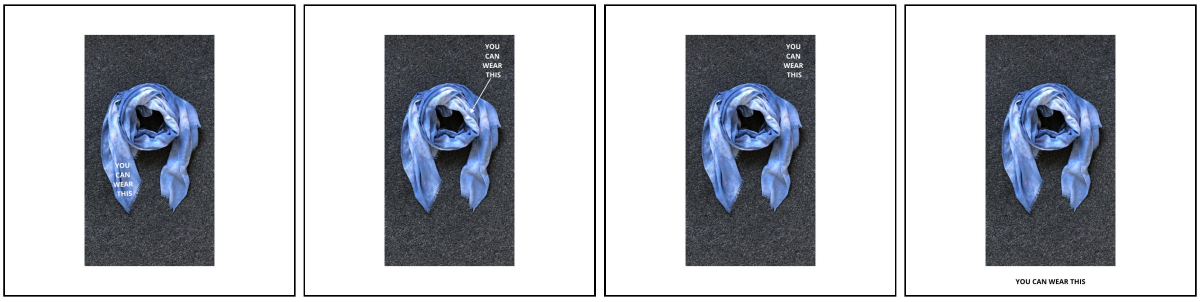
Congruence: $M = 1.03$, $SD = 0.18$

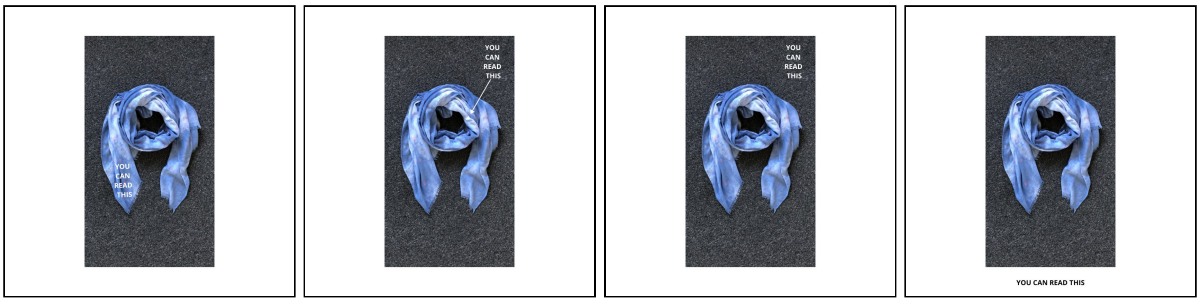
Statements

Congruent: You can wear this.

Incongruent: You can read this.

Experimental conditions





14. Socks

Pretest results

Picture validation: socks (96.67%), feet (3.33%)

Congruence: $M = 1.00$, $SD = 0.00$

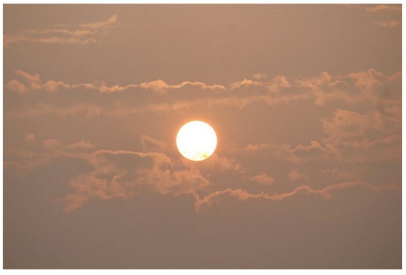
Statements

Congruent: You can put these on.

Incongruent: This can be blown away.

Experimental conditions

15. Sun



Pretest results

Picture validation: sun (96.67%), sky (3.33%)

Congruence: *M* = 1.17, *SD* = 0.59

Statements

Congruent: You can tan with this.

Incongruent: You can read time with this.

Experimental conditions

16. Tree

Pretest results

Picture validation: tree (100%)

Congruence: *M* = 1.13, *SD* = 0.43

Statements

Congruent: You can plant this.

Incongruent: You can drive this.

Experimental conditions

Appendix B

Text and Image Comprehension Survey

Start block

Multimodal Study

Thank you for taking part in this survey!

In this study, you will be seeing several images combined with text. This study aims to better understand how we comprehend multimodal messages, so before you get to this part of the experiment, we'll ask a bit about your background. Completing the survey is entirely voluntary. It will take approximately 5 minutes and will not involve any discomfort or risk beyond that normally experienced in day-to-day life.

Please complete this survey on a PC/Mac browser (we recommend Firefox or Chrome), because it will not work on a smartphone or tablet.

All the responses you enter will be stored anonymously and you will NOT be asked for any sensitive or personal information. Data will only be used for research purposes for an unlimited basis, and will be stored in a secure location for ten years. The results of this study may be published in a scholarly book or journal, or presented at professional conferences, or used for teaching purposes, but no identifiers will be used in any publication or teaching materials.

On the next page, you will be able to confirm your consent for participating for the duration of this study. If you want to withdraw at any time, or for any reason, you can do so by closing the browser. This project has been reviewed on behalf of the Ethical Review Board of Tilburg University School of Humanities and Digital Sciences and has been given approval.

Your participation helps us advance the understanding of linguistic, visual, and multimodal cognition and communication. For questions or additional information related to this study, you can contact the Principal Investigator, Neil Cohn (Tilburg University), at n.cohn@uvt.nl.

Please click the button to continue…

Informed Consent

Multimodal Study

Completing the survey is entirely voluntary. It will take approximately 5 minutes and will not involve any discomfort or risk beyond that normally experienced in day-to-day life. If you want to withdraw at any time, or for any reason, you can do so by closing the browser.
Your experimental data will be completely anonymous, so that no personally identifying information is collected.
The data will be gathered by staff and/or students at Tilburg University. We will be collecting information about true or false statements , as well as some general demographic information. The anonymised experimental data will be stored for a period of ten years, which may be reused, and may be shared in permanent, publicly accessible archives.

By consenting to this survey you confirm the following:

- I have read and understand the study information above.
- I understand that my participation is voluntary and that I am free to withdraw from the project at any time without giving any reason and without penalty.
- I understand that data collected in this project might be shared as appropriate and for publication of findings, in which case data will remain completely anonymous.
- I consent to having my data processed as described above.
- I agree to take part in the current study

IF YOU GIVE YOUR CONSENT AND WISH TO CONTINUE, SELECT THE OPTION BELOW AND CLICK CONTINUE TO START THE SURVEY....

Appendix C

Coding sheet

Item number	Scenario	InterfaceType	Congruence	Text	Image_Name	List
1	BAG	INHERENT	CONGRUENT	YOU CAN PUT SOMETHING IN THIS	BAG_A_IN_C	1
1	BAG	EMERGENT	CONGRUENT	YOU CAN PUT SOMETHING IN THIS	BAG_B_EM_C	2
1	BAG	ADJOINED	CONGRUENT	YOU CAN PUT SOMETHING IN THIS	BAG_C_AD_C	3
1	BAG	INDEPENDENT	CONGRUENT	YOU CAN PUT SOMETHING IN THIS	BAG_D_IND_C	4
1	BAG	INHERENT	INCONGRUENT	YOU CAN LOCK WITH THIS	BAG_E_IN_IC	5
1	BAG	EMERGENT	INCONGRUENT	YOU CAN LOCK WITH THIS	BAG_F_EM_IC	6
1	BAG	ADJOINED	INCONGRUENT	YOU CAN LOCK WITH THIS	BAG_G_AD_IC	7
1	BAG	INDEPENDENT	INCONGRUENT	YOU CAN LOCK WITH THIS	BAG_H_IND_IC	8
2	BED	INHERENT	CONGRUENT	YOU CAN SLEEP IN THIS	BED_A_IN_C	8
2	BED	EMERGENT	CONGRUENT	YOU CAN SLEEP IN THIS	BED_B_EM_C	1
2	BED	ADJOINED	CONGRUENT	YOU CAN SLEEP IN THIS	BED_C_AD_C	2
2	BED	INDEPENDENT	CONGRUENT	YOU CAN SLEEP IN THIS	BED_D_IND_C	3
2	BED	INHERENT	INCONGRUENT	YOU CAN PICK THIS	BED_E_IN_IC	4
2	BED	EMERGENT	INCONGRUENT	YOU CAN PICK THIS	BED_F_EM_IC	5
2	BED	ADJOINED	INCONGRUENT	YOU CAN PICK THIS	BED_G_AD_IC	6
2	BED	INDEPENDENT	INCONGRUENT	YOU CAN PICK THIS	BED_H_IND_IC	7
3	BOOK	INHERENT	CONGRUENT	YOU CAN READ THIS	BOOK_A_IN_C	7
3	BOOK	EMERGENT	CONGRUENT	YOU CAN READ THIS	BOOK_B_EM_C	8
3	BOOK	ADJOINED	CONGRUENT	YOU CAN READ THIS	BOOK_C_AD_C	1
3	BOOK	INDEPENDENT	CONGRUENT	YOU CAN READ THIS	BOOK_D_IND_C	2
3	BOOK	INHERENT	INCONGRUENT	YOU CAN WEAR THIS	BOOK_E_IN_IC	3
3	BOOK	EMERGENT	INCONGRUENT	YOU CAN WEAR THIS	BOOK_F_EM_IC	4
3	BOOK	ADJOINED	INCONGRUENT	YOU CAN WEAR THIS	BOOK_G_AD_IC	5
3	BOOK	INDEPENDENT	INCONGRUENT	YOU CAN WEAR THIS	BOOK_H_IND_IC	6
4	CAR	INHERENT	CONGRUENT	YOU CAN DRIVE THIS	CAR_A_IN_C	6
4	CAR	EMERGENT	CONGRUENT	YOU CAN DRIVE THIS	CAR_B_EM_C	7
4	CAR	ADJOINED	CONGRUENT	YOU CAN DRIVE THIS	CAR_C_AD_C	8
4	CAR	INDEPENDENT	CONGRUENT	YOU CAN DRIVE THIS	CAR_D_IND_C	1
4	CAR	INHERENT	INCONGRUENT	YOU CAN PLANT THIS	CAR_E_IN_IC	2
4	CAR	EMERGENT	INCONGRUENT	YOU CAN PLANT THIS	CAR_F_EM_IC	3
4	CAR	ADJOINED	INCONGRUENT	YOU CAN PLANT THIS	CAR_G_AD_IC	4
4	CAR	INDEPENDENT	INCONGRUENT	YOU CAN PLANT THIS	CAR_H_IND_IC	5
5	CLOCK	INHERENT	CONGRUENT	YOU CAN READ TIME WITH THIS	CLOCK_A_IN_C	5
5	CLOCK	EMERGENT	CONGRUENT	YOU CAN READ TIME WITH THIS	CLOCK_B_EM_C	6
5	CLOCK	ADJOINED	CONGRUENT	YOU CAN READ TIME WITH THIS	CLOCK_C_AD_C	7

5	CLOCK	INDEPENDENT	CONGRUENT	YOU CAN READ TIME WITH THIS	CLOCK_D_IND_C	8
5	CLOCK	INHERENT	INCONGRUENT	YOU CAN TAN WITH THIS	CLOCK_E_IN_IC	1
5	CLOCK	EMERGENT	INCONGRUENT	YOU CAN TAN WITH THIS	CLOCK_F_EM_IC	2
5	CLOCK	ADJOINED	INCONGRUENT	YOU CAN TAN WITH THIS	CLOCK_G_AD_IC	3
5	CLOCK	INDEPENDENT	INCONGRUENT	YOU CAN TAN WITH THIS	CLOCK_H_IND_IC	4
6	EGG	INHERENT	CONGRUENT	YOU CAN EAT THIS	EGG_A_IN_C	4
6	EGG	EMERGENT	CONGRUENT	YOU CAN EAT THIS	EGG_B_EM_C	5
6	EGG	ADJOINED	CONGRUENT	YOU CAN EAT THIS	EGG_C_AD_C	6
6	EGG	INDEPENDENT	CONGRUENT	YOU CAN EAT THIS	EGG_D_IND_C	7
6	EGG	INHERENT	INCONGRUENT	YOU CAN DRINK FROM THIS	EGG_E_IN_IC	8
6	EGG	EMERGENT	INCONGRUENT	YOU CAN DRINK FROM THIS	EGG_F_EM_IC	1
6	EGG	ADJOINED	INCONGRUENT	YOU CAN DRINK FROM THIS	EGG_G_AD_IC	2
6	EGG	INDEPENDENT	INCONGRUENT	YOU CAN DRINK FROM THIS	EGG_H_IND_IC	3
7	GLASS	INHERENT	CONGRUENT	YOU CAN DRINK FROM THIS	GLASS_A_IN_C	3
7	GLASS	EMERGENT	CONGRUENT	YOU CAN DRINK FROM THIS	GLASS_B_EM_C	4
7	GLASS	ADJOINED	CONGRUENT	YOU CAN DRINK FROM THIS	GLASS_C_AD_C	5
7	GLASS	INDEPENDENT	CONGRUENT	YOU CAN DRINK FROM THIS	GLASS_D_IND_C	6
7	GLASS	INHERENT	INCONGRUENT	YOU CAN EAT THIS	GLASS_E_IN_IC	7
7	GLASS	EMERGENT	INCONGRUENT	YOU CAN EAT THIS	GLASS_F_EM_IC	8
7	GLASS	ADJOINED	INCONGRUENT	YOU CAN EAT THIS	GLASS_G_AD_IC	1
7	GLASS	INDEPENDENT	INCONGRUENT	YOU CAN EAT THIS	GLASS_H_IND_IC	2
8	KEY	INHERENT	CONGRUENT	YOU CAN LOCK WITH THIS	KEY_A_IN_C	2
8	KEY	EMERGENT	CONGRUENT	YOU CAN LOCK WITH THIS	KEY_B_EM_C	3
8	KEY	ADJOINED	CONGRUENT	YOU CAN LOCK WITH THIS	KEY_C_AD_C	4
8	KEY	INDEPENDENT	CONGRUENT	YOU CAN LOCK WITH THIS	KEY_D_IND_C	5
8	KEY	INHERENT	INCONGRUENT	YOU CAN PUT SOMETHING IN THIS	KEY_E_IN_IC	6
8	KEY	EMERGENT	INCONGRUENT	YOU CAN PUT SOMETHING IN THIS	KEY_F_EM_IC	7
8	KEY	ADJOINED	INCONGRUENT	YOU CAN PUT SOMETHING IN THIS	KEY_G_AD_IC	8
8	KEY	INDEPENDENT	INCONGRUENT	YOU CAN PUT SOMETHING IN THIS	KEY_H_IND_IC	1
9	KNIFE	INHERENT	CONGRUENT	YOU CAN CUT WITH THIS	KNIFE_A_IN_C	1
9	KNIFE	EMERGENT	CONGRUENT	YOU CAN CUT WITH THIS	KNIFE_B_EM_C	2
9	KNIFE	ADJOINED	CONGRUENT	YOU CAN CUT WITH THIS	KNIFE_C_AD_C	3
9	KNIFE	INDEPENDENT	CONGRUENT	YOU CAN CUT WITH THIS	KNIFE_D_IND_C	4
9	KNIFE	INHERENT	INCONGRUENT	YOU CAN ILLUMINATE WITH THIS	KNIFE_E_IN_IC	5
9	KNIFE	EMERGENT	INCONGRUENT	YOU CAN ILLUMINATE WITH THIS	KNIFE_F_EM_IC	6
9	KNIFE	ADJOINED	INCONGRUENT	YOU CAN ILLUMINATE WITH THIS	KNIFE_G_AD_IC	7
9	KNIFE	INDEPENDENT	INCONGRUENT	YOU CAN ILLUMINATE WITH THIS	KNIFE_H_IND_IC	8
10	LAMP	INHERENT	CONGRUENT	YOU CAN ILLUMINATE WITH THIS	LAMP_A_IN_C	8
10	LAMP	EMERGENT	CONGRUENT	YOU CAN ILLUMINATE WITH THIS	LAMP_B_EM_C	1
10	LAMP	ADJOINED	CONGRUENT	YOU CAN ILLUMINATE WITH THIS	LAMP_C_AD_C	2
10	LAMP	INDEPENDENT	CONGRUENT	YOU CAN ILLUMINATE WITH THIS	LAMP_D_IND_C	3
10	LAMP	INHERENT	INCONGRUENT	YOU CAN CUT WITH THIS	LAMP_E_IN_IC	4

10	LAMP	EMERGENT	INCONGRUENT	YOU CAN CUT WITH THIS	LAMP_F_EM_IC	5
10	LAMP	ADJOINED	INCONGRUENT	YOU CAN CUT WITH THIS	LAMP_G_AD_IC	6
10	LAMP	INDEPENDENT	INCONGRUENT	YOU CAN CUT WITH THIS	LAMP_H_IND_IC	7
11	LEAF	INHERENT	CONGRUENT	THIS CAN BE BLOWN AWAY	LEAF_A_IN_C	7
11	LEAF	EMERGENT	CONGRUENT	THIS CAN BE BLOWN AWAY	LEAF_B_EM_C	8
11	LEAF	ADJOINED	CONGRUENT	THIS CAN BE BLOWN AWAY	LEAF_C_AD_C	1
11	LEAF	INDEPENDENT	CONGRUENT	THIS CAN BE BLOWN AWAY	LEAF_D_IND_C	2
11	LEAF	INHERENT	INCONGRUENT	YOU CAN PUT THESE ON	LEAF_E_IN_IC	3
11	LEAF	EMERGENT	INCONGRUENT	YOU CAN PUT THESE ON	LEAF_F_EM_IC	4
11	LEAF	ADJOINED	INCONGRUENT	YOU CAN PUT THESE ON	LEAF_G_AD_IC	5
11	LEAF	INDEPENDENT	INCONGRUENT	YOU CAN PUT THESE ON	LEAF_H_IND_IC	6
12	ROSE	INHERENT	CONGRUENT	YOU CAN PICK THIS	ROSE_A_IN_C	6
12	ROSE	EMERGENT	CONGRUENT	YOU CAN PICK THIS	ROSE_B_EM_C	7
12	ROSE	ADJOINED	CONGRUENT	YOU CAN PICK THIS	ROSE_C_AD_C	8
12	ROSE	INDEPENDENT	CONGRUENT	YOU CAN PICK THIS	ROSE_D_IND_C	1
12	ROSE	INHERENT	INCONGRUENT	YOU CAN SLEEP IN THIS	ROSE_E_IN_IC	2
12	ROSE	EMERGENT	INCONGRUENT	YOU CAN SLEEP IN THIS	ROSE_F_EM_IC	3
12	ROSE	ADJOINED	INCONGRUENT	YOU CAN SLEEP IN THIS	ROSE_G_AD_IC	4
12	ROSE	INDEPENDENT	INCONGRUENT	YOU CAN SLEEP IN THIS	ROSE_H_IND_IC	5
13	SCARF	INHERENT	CONGRUENT	YOU CAN WEAR THIS	SCAR_A_IN_C	5
13	SCARF	EMERGENT	CONGRUENT	YOU CAN WEAR THIS	SCARF_B_EM_C	6
13	SCARF	ADJOINED	CONGRUENT	YOU CAN WEAR THIS	SCARF_C_AD_C	7
13	SCARF	INDEPENDENT	CONGRUENT	YOU CAN WEAR THIS	SCARF_D_IND_C	8
13	SCARF	INHERENT	INCONGRUENT	YOU CAN READ THIS	SCARF_E_IN_IC	1
13	SCARF	EMERGENT	INCONGRUENT	YOU CAN READ THIS	SCARF_F_EM_IC	2
13	SCARF	ADJOINED	INCONGRUENT	YOU CAN READ THIS	SCARF_G_AD_IC	3
13	SCARF	INDEPENDENT	INCONGRUENT	YOU CAN READ THIS	SCARF_H_IND_IC	4
14	SOCKS	INHERENT	CONGRUENT	YOU CAN PUT THESE ON	SOCKS_A_IN_C	4
14	SOCKS	EMERGENT	CONGRUENT	YOU CAN PUT THESE ON	SOCKS_B_EM_C	5
14	SOCKS	ADJOINED	CONGRUENT	YOU CAN PUT THESE ON	SOCKS_C_AD_C	6
14	SOCKS	INDEPENDENT	CONGRUENT	YOU CAN PUT THESE ON	SOCKS_D_IND_C	7
14	SOCKS	INHERENT	INCONGRUENT	THIS CAN BE BLOWN AWAY	SOCKS_E_IN_IC	8
14	SOCKS	EMERGENT	INCONGRUENT	THIS CAN BE BLOWN AWAY	SOCKS_F_EM_IC	1
14	SOCKS	ADJOINED	INCONGRUENT	THIS CAN BE BLOWN AWAY	SOCKS_G_AD_IC	2
14	SOCKS	INDEPENDENT	INCONGRUENT	THIS CAN BE BLOWN AWAY	SOCKS_H_IND_IC	3
15	SUN	INHERENT	CONGRUENT	YOU CAN TAN WITH THIS	SUN_A_IN_C	3
15	SUN	EMERGENT	CONGRUENT	YOU CAN TAN WITH THIS	SUN_B_EM_C	4
15	SUN	ADJOINED	CONGRUENT	YOU CAN TAN WITH THIS	SUN_C_AD_C	5
15	SUN	INDEPENDENT	CONGRUENT	YOU CAN TAN WITH THIS	SUN_D_IND_C	6
15	SUN	INHERENT	INCONGRUENT	YOU CAN READ TIME WITH THIS	SUN_E_IN_IC	7
15	SUN	EMERGENT	INCONGRUENT	YOU CAN READ TIME WITH THIS	SUN_F_EM_IC	8
15	SUN	ADJOINED	INCONGRUENT	YOU CAN READ TIME WITH THIS	SUN_G_AD_IC	1

15	SUN	INDEPENDENT	INCONGRUENT	YOU CAN READ TIME WITH THIS	SUN_H_IND_IC	2
16	TREE	INHERENT	CONGRUENT	YOU CAN PLANT THIS	TREE_A_IN_C	2
16	TREE	EMERGENT	CONGRUENT	YOU CAN PLANT THIS	TREE_B_EM_C	3
16	TREE	ADJOINED	CONGRUENT	YOU CAN PLANT THIS	TREE_C_AD_C	4
16	TREE	INDEPENDENT	CONGRUENT	YOU CAN PLANT THIS	TREE_D_IND_C	5
16	TREE	INHERENT	INCONGRUENT	YOU CAN DRIVE THIS	TREE_E_IN_IC	6
16	TREE	EMERGENT	INCONGRUENT	YOU CAN DRIVE THIS	TREE_F_EM_IC	7
16	TREE	ADJOINED	INCONGRUENT	YOU CAN DRIVE THIS	TREE_G_AD_IC	8
16	TREE	INDEPENDENT	INCONGRUENT	YOU CAN DRIVE THIS	TREE_H_IND_IC	1