博士論文

The adaptive use of heuristics: Investigations of human inferential strategies in a new task structure

（ヒューリスティックの適応的な利用: 新たな課題構造における人の推論方略の検証）

白砂大

Table of contents

Chapter 1 Introduction

Herbert A. Simon, a Nobel-Prize laureate and a great figure in cognitive science and computer science, proposed an important theory about human intelligence, called *bounded rationality* (e.g., Simon, 1955, 1990). People make inferences about the real world under many constraints such as their limited knowledge and time pressure. Since it is generally difficult to carry on the best optimizing strategy based on detailed analyses, people must find the way to solve the problems approximately. Nevertheless, they can make "satisficing" (a mixed word of *sufficing* and *satisfying* by Simon; see also Gigerenzer & Goldstein, 1996), often correct, inferences even when they have several constraints described above. Such intelligence that the human system shows is referred to as bounded rationality (see also Gigerenzer & Goldstein, 1996; Hilbig & Pohl, 2009; Hoffrage & Reimer, 2004; Nickles, 2018). In order to make as plausible inferences as possible under constraints, people often use simple and intuitive inferential strategy like a rule of thumb, called *heuristics*. Previous studies have shown that heuristics sometimes do not work well (i.e., making false inferences; e.g., Tversky & Kahneman, 1973, 1974) but sometimes work well (i.e., making correct inferences; e.g., Goldstein & Gigerenzer, 2002; Schooler & Hertwig, 2005). In this regard, the *adaptive toolbox* framework has now widely been accepted: It is assumed that people "are equipped" with several inferential strategies and "select" a certain appropriate strategy among them for solving tasks (e.g., Bröder, 2003; Gigerenzer, 2001; Gigerenzer, Todd, & ABC Research Group, 1999; Scheibehenne, Rieskamp, & Wagenmakers, 2013; Todd & Gigerenzer, 2000). Then, what contributes to the adaptive nature of heuristics? One important concept is, of course, *the accuracy*: How accurate is the strategy that people used in tasks? As to the accuracy of heuristics, previous studies show that even a simple heuristic can often produce correct inferences if the structure of the heuristic matches that of the real-world environment well (i.e., *ecological rationality*; e.g., Goldstein & Gigerenzer, 2002; 2011; see also Gigerenzer & Gaissmaier, 2011). For discussing the adaptive nature of heuristics, however, another important concept, *the applicability*, should also be considered: How often can people use the strategy in tasks?

Even if a certain heuristic can reflect an environmental structure and can lead to correct inferences, it will be of no use if there are little chances to use it due to some constraints (e.g., Hilbig & Pohl, 2008; Schooler & Hertwig, 2005; Schurz & Hertwig, 2019).

In this thesis, according to the adaptive toolbox framework, I examine people's adaptive use of heuristics, focusing on a new aspect on which previous studies have not focused so far: *Task structure* (i.e., the location in a problem statement where objects are presented, and the computation that a person is required for solving the task; for details, see section 1.5). Specifically, I propose a new binary choice task (*relationships-comparison task*) and predict that people will use a new heuristic for the task (*familiarity-matching*). Then, using the new task, I investigate the strategy that people use, the strategy's accuracy, and the strategy's applicability. The outline of this thesis is as follows. Chapter 1 describes a historical overview of decision science about human intelligence related to heuristics. Furthermore, I propose a new task structure (i.e., *relationships-comparison task*) and describe the purposes of my studies for examining the adaptive use of heuristics. In Chapter 2, three inferential models are introduced (i.e., *familiarity-matching*, *familiarity heuristic*, and *knowledge-based inference*) to describe human inferences in a relationships-comparison task. In Chapter 3 (Study 1), through a behavioral experiment and model-based approaches, I investigate what strategies people often use in a relationships-comparison task. In Chapter 4 (Study 2), through a behavioral experiment, analyses of the real-world data, and computer simulations, I examine the accuracy of the heuristic (especially I focus on familiarity-matching) in terms of ecological rationality. In Chapter 5 (Study 3), through a behavioral experiment, the main findings of Study 1 are replicated. In Chapter 6 (Study 4), in the first place, Study 4a, I examine which strategies will be more adaptive in terms of the correct rate (accuracy) and applicability through analyses of the behavioral data. After that, Study 4b, the correct rate and applicability of heuristics are investigated, manipulating individuals' decision threshold (i.e., sensitivity to discriminate two objects' similarities in familiarity) through computer simulations. In Chapter 7 (Study 5), through an exploratory behavioral experiment, I examine the strategy that people use for a relationships-comparison task in a daily context, consumer choices. Chapter 8 describes the general discussion for my five studies in terms of the results of additional analyses, previous related theories, limitations, and future works. Finally, in Chapter 9, I conclude this thesis.

Before starting the main text, I summarize the definitions of important terms used in this thesis (section 1.2), and then clarify my standpoint and general focus of this thesis (section 1.1).

1.1 The important terminology in this thesis

In this thesis, I use several important terms (underlined below) according to previous studies by Gigerenzer and his colleagues. However, they did not provide clear definitions for some of these terms in their works, and the meanings of these terms often seem to be ambiguous and equivocal. These terms are often used with somewhat different meanings from the proper meanings, and therefore readers may be confused or may understand the meanings differently than I intended. So, here I provide the definitions of these important terms in this thesis.

・Information:
Contents that are seen or heard by people in the real world and that are often regarded as general knowledge (some of them are the knowledge that can be asked in a quiz related to general knowledge)

・Environment (or the real-world environment):
The real world that people see or hear through media and documents
(i.e., The real world that will shape people's subjective memory experiences, such as recognition, fluency, or familiarity, through media and documents)

・Environmental structure (or structure of environment):
The frequency of appearance of certain information in the real world

・Task structure:
A task format of a binary choice that is defined in terms of …
… the location in a problem statement where objects are presented
… the comparison or computation that a person is required for solving the task

1.2 The standpoint and the general focus of this thesis

This thesis investigated the adaptive use of heuristics according to one of the most important framework about human intelligence, *adaptive toolbox* (e.g., Gigerenzer et al., 1999). The adaptive toolbox framework explains that people have different inferential strategies and use them depending on the environment (for details, see section 1.3). So far, however, the adaptive toolbox has been examined only in one task structure of a binary choice (as described later, two alternatives were presented, such as in a *population inference task*). Although it is believed that people adaptively change their strategies, previous studies have not investigated whether people really use different strategies even in a different task structure. Thus, this thesis investigates the above issue within the adaptive toolbox framework. According to Gigerenzer and his colleagues, I focus on the followings in this thesis:

・An environmental structure similar to that used by Gigerenzer and his colleagues: Because they mainly used a population inference task, I will also use the environmental structure related to geographical features.

・A task structure different from that used by Gigerenzer and his colleagues (for details, see section 1.5)

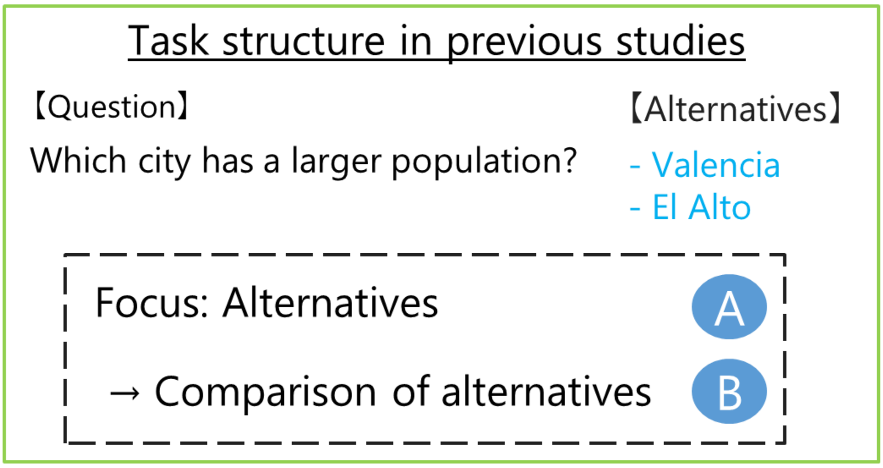
In this thesis, I conducted five studies (for the detailed purposes of these studies, see section 1.7). From Study 1 to Study 4, I used inferential tasks about general knowledge in a specific environmental structure. According to Gigerenzer and his colleagues' works, I focused on the structure of information that people see or hear through media and documents in the real world. In Study 5, on the other hand, I used the same task structure in Studies 1~4 but focused on a preferential task wherein participants were asked which item they wanted to buy and were not asked about general knowledge. I investigated whether, even if the type of task changed, people used the same heuristic as reported for a different task with the same task structure in Studies 1~4.

1.3 Historical background of heuristic studies: From "irrational" to "rational"

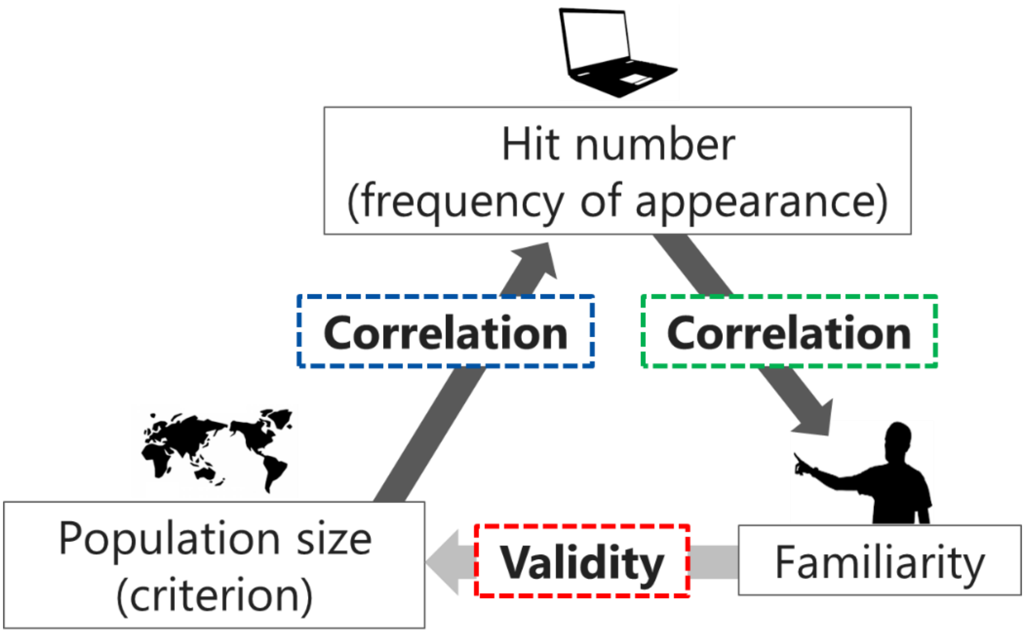
Before the 1970s, it was considered that accurate inferences were linked to logical rules (as review, Vlek, 1970). Human inferences should be made according to the principles of probability and statistics, and deviations from such principles were regarded as false inferences or biases. However, the real-world environments are often so complex and computationally intractable: There are more information than people can deal with, and it is unclear what information is or is not important. If people tried to find the optimal inferences about the real world, human minds would need a super calculator like a Laplacean Demon (e.g., Curley, MacLean, Murray, & Laybourn, 2019; Gigerenzer & Goldstein, 1996). Then, human minds often rely on simple strategies, called *heuristics*, to make inferences within their limited knowledge and computational power. The use of heuristics has been interpreted as one of the important essences of human intelligence. Traditionally, heuristics have been used as smart strategies to enhance solving problems (as review, Simon, 1990). In problem-solving studies, for example, heuristics have been regarded as important strategies for solving tasks which have large problem space (e.g., Newell, Shaw, & Simon, 1959; Simon & Newell, 1971). Some computational models proposed in such studies have been the basis of today's human cognitive models (e.g., ACT-R, which is described in Study 4 in this thesis) (e.g., Anderson et al., 2004).

In the 1970s~1980s, however, it was found that heuristics were so simple that people often made false inferences about the real world. For example, people often use information based on availability (e.g., information that is easily available) as their inference cues. However, consider the following question: In an English text, is it more likely that the word starts with a K, or that K is its third letter? Actually, the answer is the latter alternative (i.e., K is its third letter), but many people choose the former alternative (i.e., start with K). This is because the former cases are more available (i.e., easy to retrieve from their knowledge, e.g., Know, Kind, Knight, etc.) than the latter cases (e.g., Tversky & Kahneman, 1974). Researchers at that time concluded that human inferences were systematically biased by using heuristics, and that heuristics were "irrational" inferential strategies (e.g., Kahneman & Tversky, 1972; Tversky & Kahneman, 1973, 1974, 1981, 1983; as review, Chase, Hertwig, & Gigerenzer, 1998).

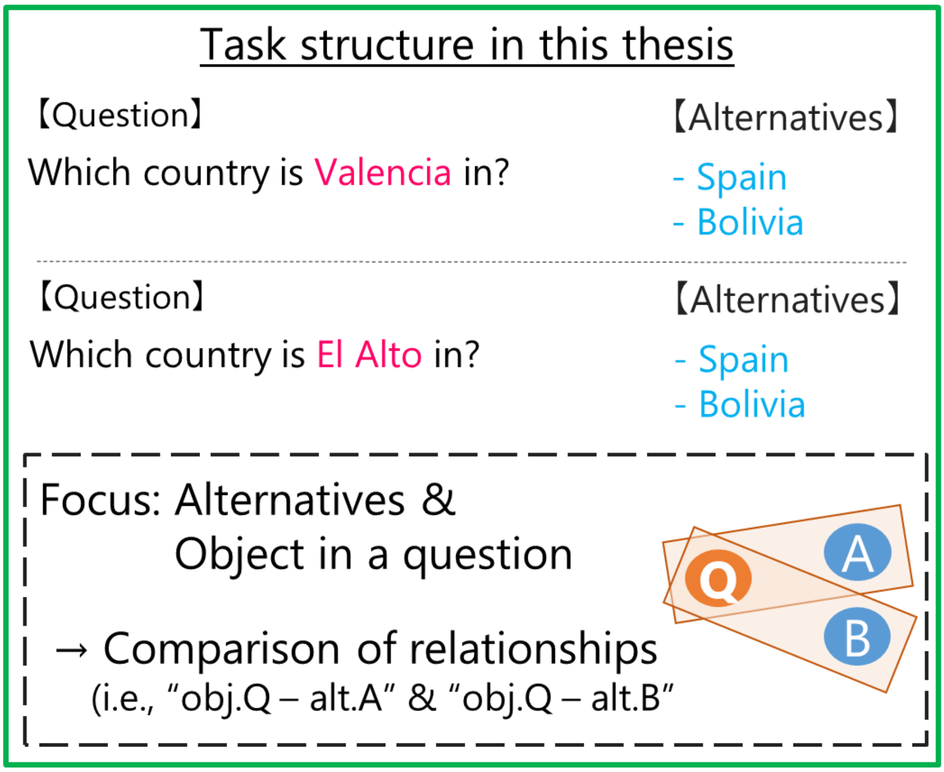
After the 1990s, in contrast, the adaptive nature of heuristics have been reported (e.g., Gigerenzer et al., 1999; Goldstein & Gigerenzer, 1999; Todd & Gigerenzer, 2000). Gigerenzer and his colleagues have been assumed that people have various strategies and employ them for solving a given problem adaptively. Such a framework is called the *adaptive toolbox* (e.g., Bröder, 2003; Gigerenzer et al., 1999; Kruglanski & Gigerenzer, 2011; Mohnert, Pachur, & Lieder, 2019; Newell, 2005). More specifically, even if people do not or cannot search inferential cues thoroughly due to their cognitive constraints for solving tasks, they can often make inferences accurately by exploiting an environmental structure and by using a simple heuristic that will match it. The theory that captures the adaptive nature of heuristics in terms of minds and environments[i] is known as *ecological rationality* (e.g., Gigerenzer, 2001; 2008; Goldstein & Gigerenzer, 2002; Todd & Gigerenzer, 2012; see also Chater et al., 2018; Otworowska, Blokpoel, Sweers, Wareham, & van Rooij, 2018; Rieskamp & Reimer, 2007; Schurz & Hertwig, 2019). In studies on human inferences[ii], a famous task for investigating ecological rationality of heuristics is a *population inference task* (Fig. **1** (A)) (e.g., Goldstein & Gigerenzer, 2002; Hilbig, Erdfelder, & Pohl, 2011; Hilbig & Pohl, 2008; Schooler & Hertwig, 2005). In this task, people are presented two alternatives (two city names) and are asked to consider a binary choice question: "Which city has a larger population, Valencia or El Alto?" If people do not have some specific knowledge (e.g., whether the presented cities have famous soccer teams or not), they will rely on heuristics based on their subjective memory experiences such as which city they can recognize (*recognition heuristic*; e.g., Gigerenzer & Brighton, 2009; Gigerenzer & Goldstein, 2011; Goldstein & Gigerenzer, 2002; Pachur, Todd, Gigerenzer, Schooler, & Goldstein, 2011), which city they can remember more quickly (*fluency heuristic*; e.g., Hertwig, Herzog, Schooler, & Reimer, 2008; Schooler & Hertwig, 2005), or which city they are more familiar with (*familiarity heuristic*; e.g., Honda, Abe, Matsuka, & Yamagishi, 2011; Honda, Matsuka, & Ueda, 2017; Xu, González-Vallejo, Weinhardt, Chimeli, & Karadogan, 2018).[iii] Many people consider that the more familiar alternative has the higher value to a criterion (e.g., population size). In this example, people tend to choose "Valencia" because, for many people, Valencia is more familiar than "El Alto." Interestingly, such a simple, familiarity-based heuristic can make correct inferences in many cases. The reason why such a heuristic works well in this task is that the structure of heuristic (i.e., the way people use their memory such as familiarity) effectively captures the structure of an environment: Larger cities tend to appear in the real world (e.g., mentioned in media) more frequently than smaller cities, and then are likely to become familiar to many people. More specifically, the frequency of appearance of objects in the real world is highly correlated with people's familiarity (a right arrow in Fig. **1** (B)), and the criterion of objects (e.g., population size) is also highly correlated with the frequency of appearance (a left arrow in Fig. **1** (B)). Thus, a heuristic of choosing the more familiar alternative (i.e., familiarity heuristic) has ecological rationality (a below arrow in Fig. **1** (B)) (e.g., Brighton, 2020; Gigerenzer, 2001; Gigerenzer & Goldstein, 1996; Goldstein & Gigerenzer, 2002; Honda et al., 2017). By focusing on such an environmental structure, researchers have clarified that heuristics do not always produce biased inferences and can become accurate inferential strategies. Such cognitive systems to make correct inferences under cognitive constraints are regarded as the "adaptivity" or "adaptive nature" of human inferences. For discussing the adaptive toolbox framework, researchers have paid attention to the importance of considering an interaction between the human minds and the real-world environments.


Task structure in previous studies
【Question】
Which city has a larger population?
【Alternatives】
- Valencia
- El Alto
Focus: Alternatives
A
→ Comparison of alternatives
B

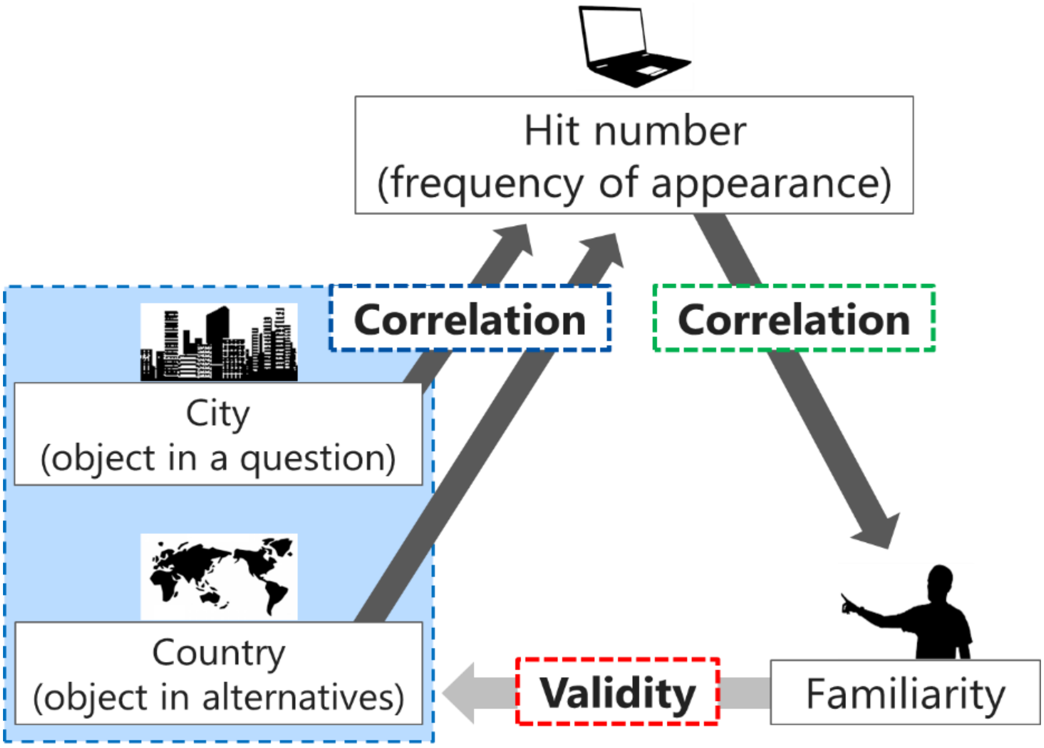
(A)


Hit number
(frequency of appearance)
Correlation
Correlation
Population size
(criterion)
Validity
Familiarity

(B)



(C)



(D)

Fig. 1 Task structures and environmental structures for the previous population inference task ((A) and (B) [adapted from Goldstein and Gigerenzer (2002) Figure 1], respectively) and for the current relationships-comparison task ((C) and (D), respectively). These two tasks have different task structures. It is predicted that if a task structure differs, then the environmental structure that people may exploit and the way that people use their familiarity will also differ, and therefore an adaptive heuristic for the task will also differ.

1.4 Recent studies: Simon's scissors metaphor and importance of studying heuristics

Based on these findings, it has been discussed that researchers should investigate human inferences considering not only the aspects of "cognition" (i.e., minds of individuals) but also those of "context" (i.e., environments in which the individuals are experiencing and are embedded)[iv]. To emphasize the need of considering these two aspects, Simon (1990) proposed a metaphor that effective inferences were generated when context and cognition fitted together well like the blades of a pair of scissors (i.e., *Simon's scissors*): One blade represents the way information is structured in real-world environments ("context" blade) and the other represents human inferences ("cognition" blade) (see also Gigerenzer et al., 1999; Hoffmann, Bettina, & Rieskamp, 2019; Kozyreva & Hertwig, 2019; Lockton, 2012; Todd & Brighton, 2016). In solving problems under constraints, people cannot always use every resource and will rely on simple heuristics (e.g., Gigerenzer et al., 1999; Newell & Bröder, 2008; Simon, 1955; van Rooij, 2008). Nevertheless, as shown in the above, when the structure of heuristics matches well to the structure of environments, people can often make correct inferences by using heuristics. So, in order to understand the adaptive use of heuristics, it is necessary to pay attention to context aspects as well as cognition aspects.

Especially in recent years, the theories about the adaptive use of heuristics such as bounded rationality and adaptive toolbox have attracted attentions of many researchers in various fields. In cognitive science, researchers can obtain implications of how people will search information, make decisions, and take actions in today's complex real-world environments (e.g., Phillips, Hertwig, Kareev, & Avrahami, 2014; Rahnev, 2020; Ratcliff, Smith, Brown, & McKoon, 2016; Sang, Todd, Goldstone, & Hills, 2020). Beyond cognitive science, in business researches, theoretical and empirical findings about simple heuristics reported in cognitive science have been applied to consider how people should make portfolio decisions and financial decisions of large scale projects in organizations (e.g., Durbach, Algorta, Kabongo, Katsikopoulos, & Şimşek, 2020; Forbes, Hudson, Skerratt, & Soufian, 2015; Ghezzi, 2020; Long, Fernbach, & De Langhe, 2018; Loock & Hinnen, 2015). Furthermore, in artificial intelligence (AI) fields, to understand and develop human-like intelligence, the interesting facts that human intelligence has inverse features to AI have drawn people's attentions: AI has been growing up by implementing enormous computational power and learning huge data, while human intelligence has limited computational power and knowledge, but both of them can often make correct inferences (e.g., Griffiths, 2020; Lake, Ullman, Tenenbaum, & Gershman, 2016; Thompson, Greenewald, Lee, & Manso, 2020). Because heuristics used by people in bounded situations is closely related to many fields, it is very important to study the adaptive use of heuristics (recently, the theory of *resource rationality* has been proposed in that people use their limited resources rationally; e.g., Lieder & Griffiths, 2020; Rich, Blokpoel, Haan, & van Rooij, 2020).

1.5 Proposed task and heuristic: Relationships-comparison task and familiarity-matching

The theories described above suggest that environments can make people smart (Todd & Gigerenzer, 2007). If so, people will show different adaptive inferences in different contexts. In this thesis, I aimed to examine human inferences according to the adaptive toolbox framework. Note that, the scopes of Gigerenzer and his colleagues' works were specific environments, as explained earlier. That is, they limited their experimental materials within the information which is seen or heard through media in the real world and is regarded as general knowledge, such as the population of cities. Because I investigate the adaptive use of heuristics according to their adaptive toolbox framework, I also focus on such a specific environmental structure as in their works (from Study 1 to Study 4 in this thesis).

Specifically, I focused on a *task structure* of a binary choice as a new context aspect that have not been investigated previously. In this thesis, I capture the task structure in terms of two points. The first point is the location in a problem statement where objects are presented. In previous binary choice tasks, two objects (A and B) were presented as alternatives (Fig. **1** (A); e.g., population inference task, as described earlier). The second point is the comparison or computation that a person is required for solving the task. In previous tasks, a person was required to compare the presented two alternatives (i.e., comparison of two objects: "A" and "B").

As a new inferential task, on the other hand, consider the following question: "Which country is Valencia in, Spain or Bolivia?" In the new task, one object (Q) is presented in a problem statement along with two objects (A and B) in alternatives (Fig. **1** (C)); and a person will have to consider not only alternatives A and B but also an object Q (i.e., comparison of two dyad relationships: "Q and A" and "Q and B"; not a simple comparison of "A" and "B").

Then, in this new task structure, do people use different strategies from those reported in previous tasks? If people have specific knowledge about Valencia, Spain, and Bolivia but do not know the correct answer, they will make inferences based on their knowledge (e.g., attributes that are relevant to the question, such as "which regions do city Q, country A, and country B belong to?"). However, when they do not have specific knowledge, they will rely on heuristics based on an available cue such as familiarity. For example, people will infer, "I am familiar with Valencia, and I am familiar with Spain, but I am not familiar with Bolivia. Valencia is more similar to Spain than to Bolivia in terms of familiarity. Therefore, Valencia should be in Spain!" In addition, consider another question: "Which country is El Alto in, Spain or Bolivia?" and assume that people are not familiar with El Alto. People will infer, "El Alto is more similar to Bolivia than to Spain in terms of familiarity. Therefore, El Alto should be in Bolivia!" As shown in these examples, when using familiarity-based heuristics in this new task structure, people will make inferences based on a similarity in familiarity. That is, people first calculate the degree of a similarity between "Valencia and Spain" as well as that between "Valencia and Bolivia" in terms of familiarity (i.e., calculating a similarity in familiarity between "Q" and "A," and calculating that between "Q" and "B"). People will next compare these two similarities in terms of familiarity (i.e., comparing the calculated similarity between "Q and A" and that between "Q and B"), and then choose Spain or Bolivia whose familiarity is more similar to that of Valencia. In this new task structure, I expect that an environmental structure that may be exploited for solving tasks will also differ from that reported in previous tasks. I will discuss the environmental structure (and ecological rationality) in the next section because this issue is related to *the accuracy* of heuristics.

In examining the adaptive toolbox of human inferences, no previous studies have directly focused on a task structure in terms of the location where objects are presented and the computation that is required for solving tasks. However, as McCloy, Beaman, and Smith (2008) implied, people's inferential performances for solving tasks would become different from those observed in previous studies when the number of alternatives increased. So, it is possible that people's heuristics will differ when a task structure differs from that addressed in previous studies. In order to obtain further understandings of people's adaptive use of heuristics, I propose this new task structure, calling it a *relationships-comparison task*. Furthermore, I also predict a new familiarity-based heuristic which may be a useful strategy for a relationships-comparison task (i.e., if the familiarity of object Q is more similar to that of alternative A than to that of alternative B, then people choose alternative A; for the detailed definition, see section 2.1), calling it as *familiarity-matching*.

1.6 The adaptive use of heuristics: The accuracy and the applicability

One may claim that it is quite natural and obvious that heuristics that people use will change if a task structure changes. It may be obvious that, if the structures of tasks are different from each other, the strategies for solving the tasks are also different. In examining human intelligence in the adaptive toolbox framework, however, it will be important to investigate whether people "adaptively" change their strategies: Do people really use a heuristic which will be adaptive for a relationships-comparison task? In particular, the adaptive use of heuristics should be evaluated in terms of the following two concepts.

One concept is *the accuracy*: How much can people make correct inferences by using a certain heuristic? To address this issue, I will focus on ecological rationality of a heuristic (especially, familiarity-matching) as described earlier. In the previous population inference task, people often choose the more familiar alternative (i.e., familiarity heuristic) and this heuristic has ecological rationality (see Fig. **1** (A) and (B)). In the new relationships-comparison task, on the other hand, since a task structure differs from the population inference task (Fig. **1** (C)), the structure of an environment that people will exploit and that of heuristic (i.e., the way people use familiarity) will also differ. Therefore, a different heuristic will be ecologically rational. Specifically, one correlation of the environmental structure will be identical to the previous tasks: The frequency of appearance of objects is highly correlated with familiarity (a right arrow in Fig. **1** (D)). However, another correlation will differ: The frequency of appearance of objects in a question will be highly correlated with that of objects in alternatives. In the above case, cities which are frequently seen or heard through media will be in countries which are also frequently seen or heard (two left arrows in Fig. **1** (D), shown in a blue-lighted box). Thus, it is expected that choosing the alternative whose familiarity is more similar to that of an object in a question (i.e., familiarity-matching) has ecological rationality, and people will often use it in a relationships-comparison task.

However, the accuracy alone is not enough to evaluate the adaptive use of heuristics. Another concept, *the applicability* is also important: How often can people use a certain heuristic in tasks? Even if a heuristic is highly accurate in tasks, it is not useful for solving them if there are little chances for people to use it. For example, recognition heuristic (i.e., choosing the recognized alternative in a binary choice) is regarded as an ecologically rational heuristic (e.g., Goldstein & Gigerenzer, 2002), but the chances to use it are limited: Recognition heuristic is applicable only when one alternative can be recognized and the other cannot, and is not applicable when both two alternatives are recognized or unrecognized (e.g., Marewski, Gaissmaier, Schooler, Goldstein, & Gigerenzer, 2009; Schooler & Hertwig, 2005; Schooler & Hertwig, 2005; Schurz & Hertwig, 2019). In the case of a relationships-comparison task, if people cannot discriminate between similarities in familiarity (i.e., they feel that the similarity of “Q and A” is almost identical to that of “Q and B”), then they will not be able to use familiarity-matching. Even if familiarity-matching is an ecologically rational heuristic but is not applicable in a relationships-comparison task, then it will not be regarded as a useful strategy.

In sum, my general question is: Do people adaptively use a certain inferential strategy which will have the higher accuracy and applicability? By analyzing these two concepts in a new task structure, the relationships-comparison task, I expect to reveal, not only whether people will simply change inferential strategies, but also whether they used a more accurate and more applicable strategy, according to the adaptive toolbox framework.

1.7 Purpose of each study

To obtain a deeper understanding of the adaptive use of heuristics according to the adaptive toolbox framework, I focused on a task structure as a new "context" blade in Simon's scissors. Through five studies, I investigated the following three issues: (1) The strategy that people use, (2) The accuracy, and (3) The applicability, in a relationships-comparison task.

First, as to the strategy that people use, my main purpose was to identify people's inferential strategies in a new task structure. I conducted a behavioral experiment to examine which strategy would often be used in a relationships-comparison task by model-based analyses (Study 1). I also conducted an identical experiment in order to replicate and to confirm the main results of Study 1 (Study 3). Furthermore, I conducted another behavioral experiment in order to examine the heuristic that people might use in a preferential, not inferential, context (Study 5).

Second, as to the accuracy, my main purpose was to clarify whether the heuristic that was often used in a relationships-comparison task would match an environmental structure, and whether the heuristic could lead to correct inferences (i.e., ecological rationality). I conducted a behavioral experiment, analyses of real-world data, and computer simulations to examine ecological rationality (Study 2).

Third, as to the applicability, my main purpose was to examine performances of strategies in terms of how applicable (as well as how accurate) each strategy would be in a relationships-comparison task. I conducted analyses of behavioral data (Study 4a); and then conducted computer simulations to manipulate individuals' cognitive constraints, *decision threshold* (i.e., the sensitivity to discriminate between similarities in familiarity) (Study 4b).

Chapter 2 Inferential models for the relationships-comparison task

First, it is not clear what strategies people will use (especially, whether they use familiarity-matching) in a relationships-comparison task because this task is a completely new task. So, I should clarify their inferential strategies.

In this thesis, I adopt the model-based approaches. I construct inferential models to describe familiarity-based heuristics and knowledge-based inferences, and then fit these models to behavioral data. Generally, it is difficult to explicitly identify the strategy that people really used in tasks. Then, previous studies have pointed out the importance of model-based approaches in examining human inferences (e.g., Gigerenzer & Brighton, 2009; Honda et al., 2017; Jenny, Rieskamp, & Nilsson, 2014; Pachur & Aebi-Forrer, 2013). In order to rigorously estimate people's inferential strategies, researchers should compare the models in terms of how well each model can explain people's inferential patterns. Note that, because researchers often have to discuss behaviors solely based on the relationships between participants' inputs and outputs (tasks and responses, respectively), the type of the models is often assumed as "as-if" model: People do not always make inferences in the exact same way as the model that researchers constructed, and the inferential strategies that people were considered to use are identified by researchers' estimation based on the results of the model fitting (e.g., Gigerenzer, 2019; Hoffrage & Reimer, 2004b; van Rooij, Wright, Kwisthout, & Wareham, 2018). In Marr's (1982) words about the levels of analyses, the models provide explanations about human inferences at a *computational level* (i.e., explanations of what output is produced from a given input), not at an *algorithmic level* (i.e., explanations of how an output is produced from a given input) (e.g., Blokpoel, Kwisthout, van der Weide, Wareham, & van Rooij, 2013; Marr, 1982; Rich et al., 2019; van Rooij & Baggio, 2021).

In the new experimental task, *relationships-comparison task*, the following format was used: "In which country is city Q, country A or country B?"[v] In this format, there are three objects: Country A, country B (in alternatives) and city Q (in a problem statement) (Fig. **1** (C)), and people may use some heuristic strategy (e.g., familiarity-based inference) or may make inferences based on their knowledge (e.g., attributes of each object that are relevant to the question; such as "which regions do city Q, country A, and country B belong to?").

In this thesis, my general aim was to identify the strategy that people use, and to clarify its accuracy and applicability for relationships-comparison tasks. To describe human inferences, three inferential models were introduced: Two heuristic models (*familiarity-matching* [FM] and *familiarity heuristic* [FH]) and one knowledge-based inference model (*knowledge-based inference* [KI]). Hereafter, I note the familiarity of an object in a question, that of alternative A, and that of alternative B, as "FamQ," "FamA," and "FamB," respectively.

2.1 Familiarity-matching (FM)

Familiarity-matching (FM) model suggests that people make inferences based on a similarity in familiarity. It is defined by the numerical distance of the familiarity ratings provided in the measurement of the familiarity task (for the detailed experimental procedure, see Study 1) between an object in a question and each of two alternatives. The assumption of FM model is as follows: A person chooses the alternative whose familiarity is more similar to that of the object in the question. FM is operationally assumed as follows: if FamQ is more similar to FamA than FamB (i.e., "$|\mathrm{FamQ} - \mathrm{FamA}| < |\mathrm{FamQ} - \mathrm{FamB}|$"), then the person chooses alternative A (Fig. 2). In my heuristic models, I introduce a concept of a *decision threshold*, which determines whether s/he can discriminate the difference in the similarity and can apply FM (e.g., Honda et al., 2017). Here I assume if the values of $|\mathrm{FamQ} - \mathrm{FamA}|$ and $|\mathrm{FamQ} - \mathrm{FamB}|$ are very similar to each other, the person will not discriminate between them and will guess (i.e., choose one out of the two alternatives randomly).

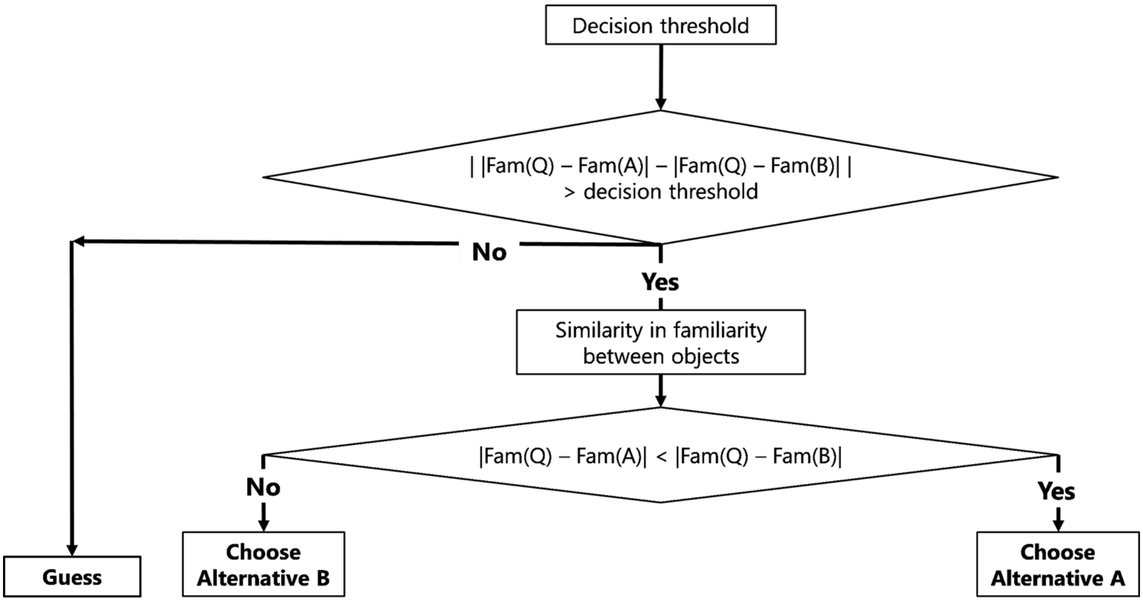


Fig. 2 Flowchart of familiarity-matching (FM) algorithm.

2.2 Familiarity heuristic (FH)

As a competitive model to FM, another familiarity-based inference model is constructed. Previous empirical and modeling studies showed that people often made inferences based on the familiarity of alternatives in binary choice questions (e.g., Honda et al., 2011; Xu et al., 2018). This type of heuristic is called *familiarity heuristic* (FH) (Honda et al., 2017). Although the familiarity heuristic in its original form assumes that people constantly choose the more familiar alternative, in this thesis the assumptions was modified to make it applicable to a relationships-comparison task. Specifically, FH model assumes that a person first consider whether an object in a question is familiar or unfamiliar. Then, if it is familiar (unfamiliar), they choose the more familiar (unfamiliar) alternative. FH is operationally assumed as follows. First, consider whether FamQ is above (below) the median of FamQs (i.e., the median of all of the person's ratings for object Q; hereafter "medianFamQs") and, if FamQ is above (below) the medianFamQs, then the person chooses the more familiar (unfamiliar) alternative (Fig. 3). As with FM, a decision threshold is introduced in making inferences. FH can be applied when a person can discriminate the difference in familiarity between FamA and FamB, and when they feel that FamQ is familiar or unfamiliar "enough" (i.e., the difference between FamQ and the medianFamQs is above the threshold). Otherwise, the person guesses (i.e., chooses one of the two alternatives randomly)[vi].

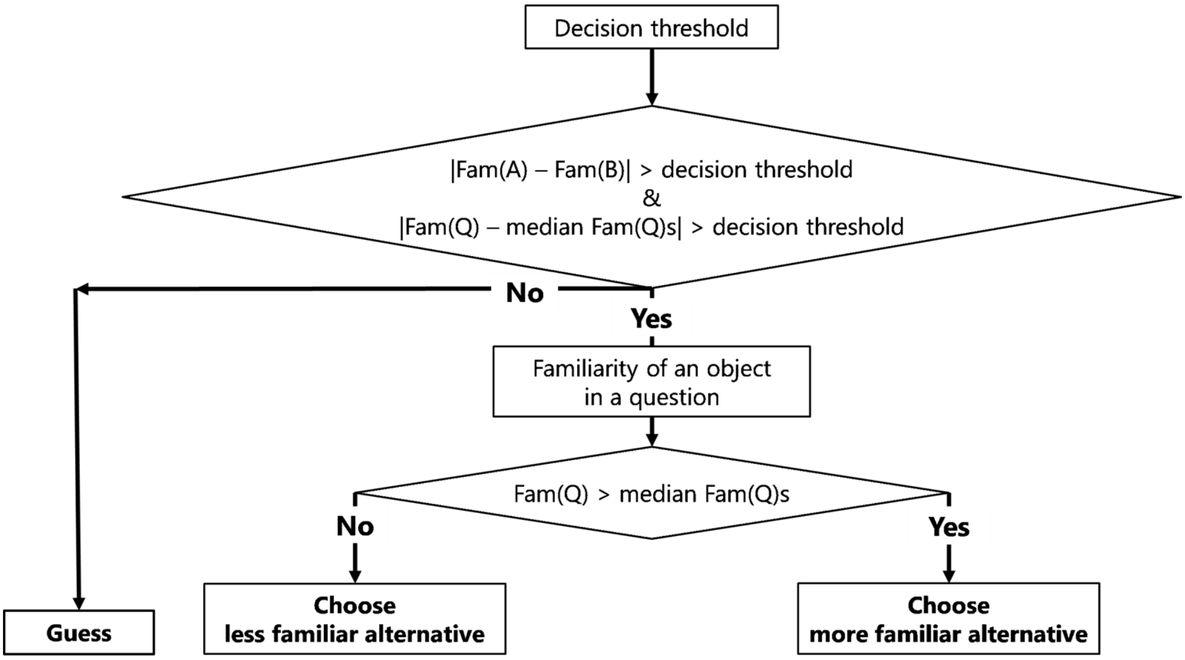


Fig. 3 Flowchart of familiarity heuristic (FH) algorithm.

2.3 Knowledge-based inference (KI)

People may make inferences using available further information or knowledge (e.g., Hilbig & Pohl, 2008, 2009; Pachur & Hertwig, 2006; Richter & Späth, 2006) such as attributes of an object, particularly when they found a task subjectively easy. For example, when people see the name "Valencia," they may remember that "Valencia is in Spain" or "Valencia is a city in Spanish-speaking countries." Then, to predict that people use such specific knowledge about objects when making inferences, a *knowledge-based inference* (KI) model is constructed. The knowledge that people use in making inferences will be different for different people. Previous studies (e.g., Castela, Kellen, Erdfelder, & Hilbig, 2014; Erdfelder et al., 2009) have shown that some people used their available knowledge in inference tasks. However, it will be difficult to identify the specific knowledge that each person used to make inferences since the person had different knowledge. Then, in this thesis, the KI model is constructed by integrating some possible inferential models, based on the following assumptions. A person uses the following four attributes that will be relevant for inferring the question: "*country*," "*region*," "*language*," and "*religion*." In KI model, inferential strategies are operationally assumed to be represented by the six *Lexicographic* models (LEX; e.g., Hoffrage & Reimer, 2004; Payne, Bettman, & Johnson, 1993) and two *Tally* models (TAL; e.g., Gigerenzer & Goldstein, 1996; Lee, Gluck, & Walsh, 2019; Parpart, Jones, & Love, 2018). LEX predicts that a person checks the content of an object's attributes (e.g., "What is the *language* in city Q, country A, or country B?"). When the attribute of city Q matches that of country A but does not match that of country B, the person infers that city Q is in country A. When an attribute for city Q matches those for both countries A and B or does not match either, a person continues to check the next attribute (Fig. 4). As for the order of checking attributes, the following procedure is assumed. Since my inferential question is about "country," a person will first think which country the city Q is located in. So, at first, it is assumed that the person refers to the "country" attribute for city Q. Then, if the person does not know the country of city Q, s/he refers to one of the other three attributes until s/he can discriminate between two alternatives. Out of the all patterns of possible orders (i.e., the order that can explain her/his inferential patterns best among 3! = 6 patterns), the best-fitted order is used as her/his LEX model.

TAL predicts that the person considers the number of attributes of an object in a question that are the same as those for alternatives. TAL is operationally defined as follows (Fig. 5). At first, the person refers to the "country" attribute. If s/he thinks that city Q is in one of the two countries, s/he directly chooses the country (i.e., identical to LEX). Otherwise, the person refers to the other three attributes. Here, two TAL models (e.g., Czerlinski, Gigerenzer, & Goldstein, 1999; Dawes, 1979; Parpart et al., 2018) is introduced. TAL 1 considers the one pattern: "match". If an attribute of city Q matches that of country A, then the value of country A is added "+ 1." If it does not match or the person does not know about the attribute, then s/he adds "+ 0." That is, TAL 1 calculates the number of city Q's attributes that are the same as those for one country. In contrast, TAL 2 considers the three patterns: "match," "do not match," and "do not know (i.e., the person has no knowledge)." For each pattern, the person adds "+ 1," "−1," or "+ 0," respectively, to the value of each country. In both TAL 1 and TAL 2, the process is applied for all three attributes. After calculating values for all attributes, the person chooses the country that has the higher value. If values between two countries are equal, then s/he "guesses." Although TALs 1 and 2 generally make analogous predictions, I show an example of the operation of TALs 1 and 2 that will make a different prediction. A person does not know the country of city Q and has her/his pieces of knowledge about three attributes: city Q and country A "match," "do not match," and "do not know," respectively; while city Q and country B "match," "do not match," and "do not match," respectively. In this case, TAL 1 calculates both A's value and B's value as "+ 1 (= 1 + 0 + 0)," and therefore predicts that the person guesses randomly. In contrast, TAL 2 calculates A's value as "0 (= 1 – 1 + 0)" and B's value as "−1 (= 1 – 1 – 1)," and therefore predicts that the person chooses country A.

Finally, LEX, TAL1, and TAL2 are compared; and then the best-fitting model are regarded as the KI model.

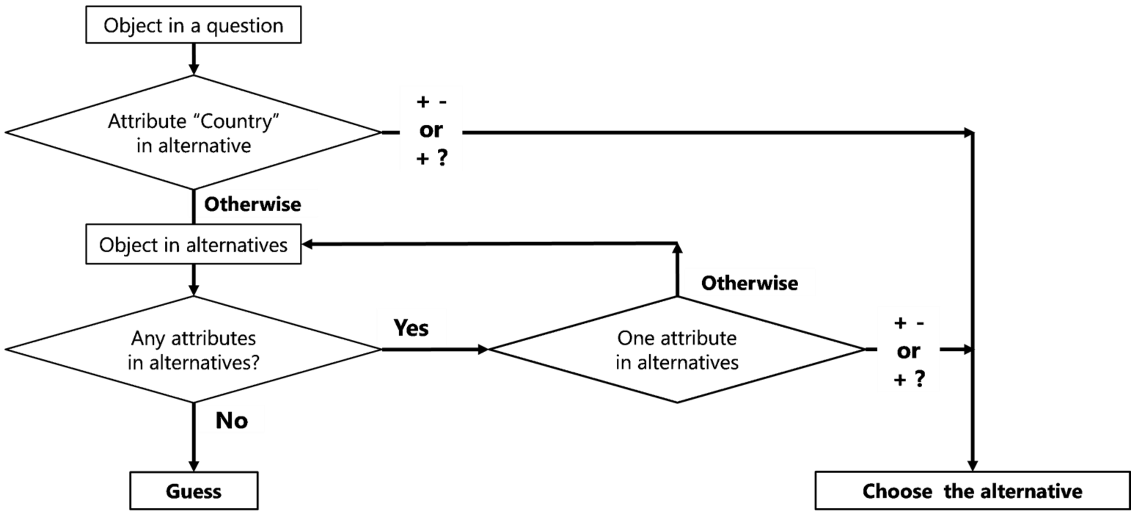


Fig. 4 Flowchart of lexicographic (LEX) algorithm in knowledge-based inference (KI) model.

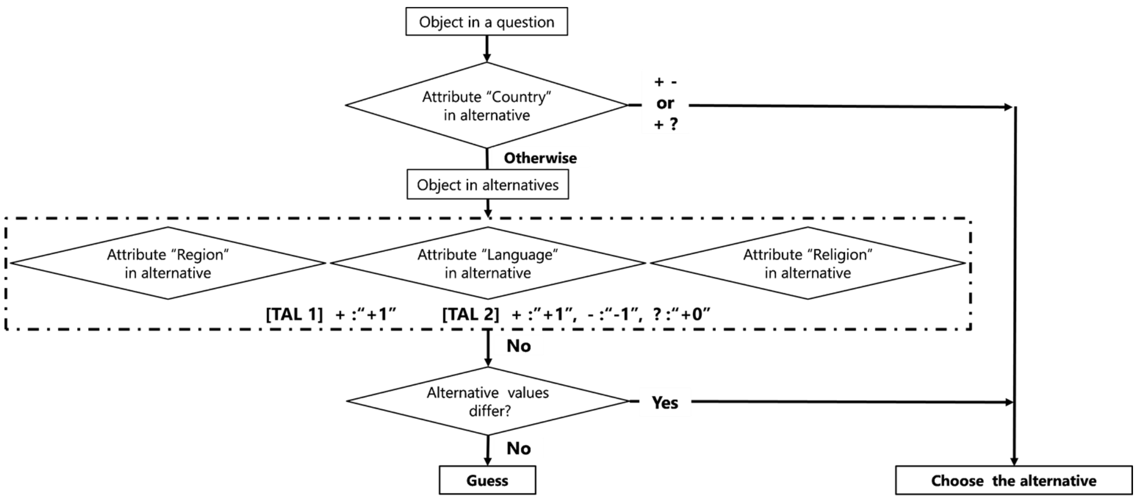


Fig. 5 Flowchart of Tally (TAL) algorithm in knowledge-based inference (KI) model.

Chapter 3 Study 1: Investigation of people's inferential strategies in a relationships-comparison task (Behavioral experiment)

In this thesis, a task structure of binary choice was focused on as a new "context" blade in Simon's scissors. The purpose of Study 1 was to identify, based on a model-based approach, what inferential strategies people use for a relationships-comparison task. I expected that people would use familiarity-matching (FM) in a relationships-comparison task because the structure of FM would match to that of the real-world environment. However, a relationships-comparison task was a newly proposed task in this thesis, therefore it was unclear which strategy people were likely to use for this new task. So, first, the inferential strategy that people really used had to be investigated.

In examining the use of strategies, the *attribute substitution* framework (Honda et al., 2017; Kahneman & Frederick, 2002, 2005) is introduced. Generally, people do not always use memory-based heuristics. A recent study focused on attribute substitution in inferences and showed that people tended to use heuristics for difficult questions more often than for easy questions (e.g., Honda et al., 2017; see Fig. 6). According to Kahneman and Frederick (2005), the attribute substitution is defined as follows: If a judgmental object (target attribute) is less readily assessed than a seemingly plausible aspect that can be easily assessed (heuristic attribute), then individuals substitute the heuristic attribute for the target attribute and make inferences based on the heuristic attribute. Honda et al. (2017) predicted that when people did not feel subjective difficulty in solving a task (e.g., they felt confident about the answer to the question), they would choose one alternative based on their available knowledge. In contrast, when people felt subjective difficulty (i.e., they did not feel confident in the answer), they would make inferences relying on heuristics after thinking that their knowledge would be unavailable for the task.

Similar to Honda et al. (2017), I compared the data of difficult questions with the data of easy questions in terms of attribute substitution framework. I predicted that, in a relationships-comparison task, people would often use a heuristic (especially, familiarity-matching) in difficult questions while would often use their knowledge in easy questions.

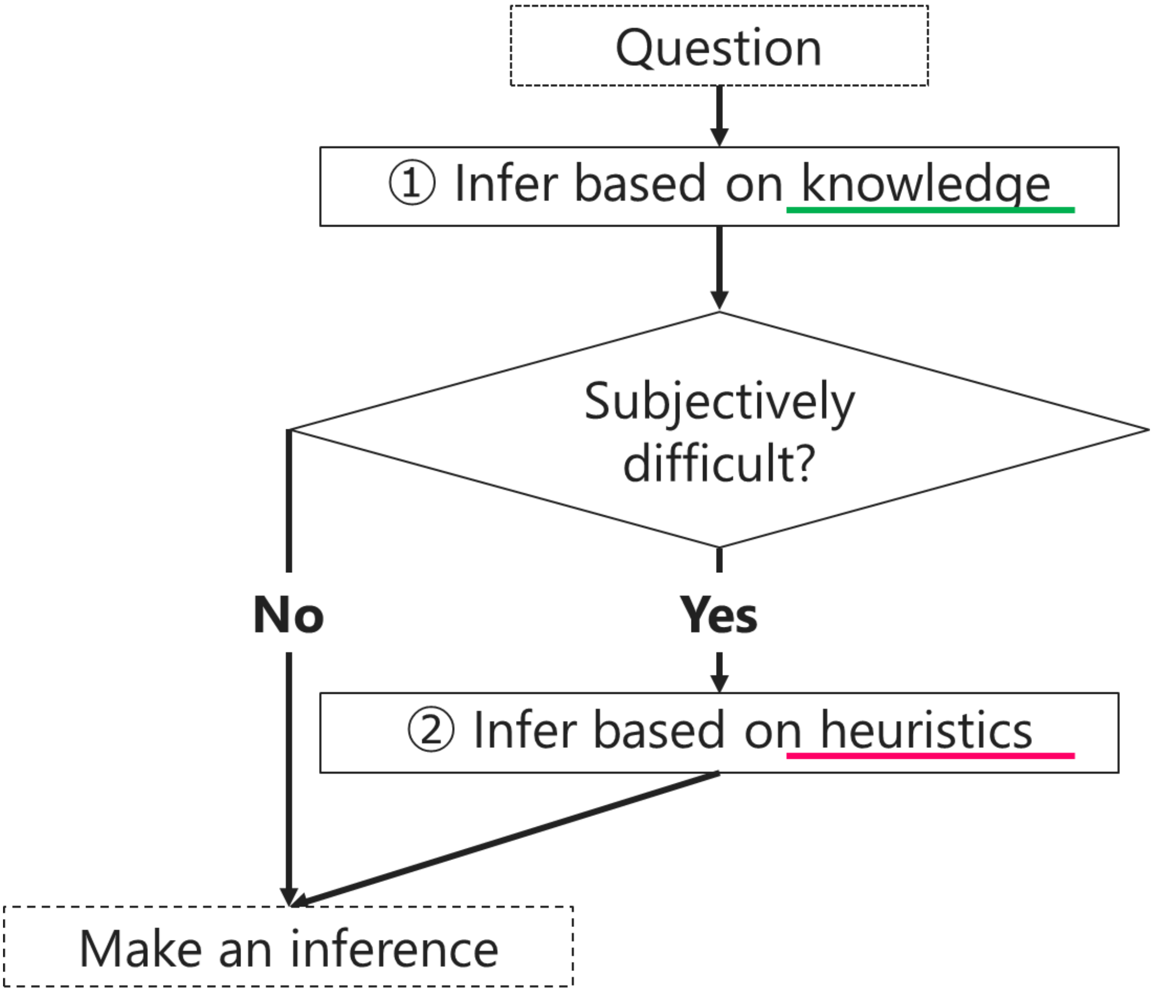


Fig. 6 Flowchart of attribute substitution framework (adapted from Honda et al., 2017). In making inferences, people first try to solve the question by using their knowledge about it (green line; e.g., knowing the correct answer or having further information). If they feel difficulty for solving the question (i.e., cannot use knowledge), then they try to solve it by relying on heuristics (pink line; e.g., using familiarity with objects).

The main procedure of Study 1 was as follows. In a behavioral experiment, participants were asked to answer the binary choice task, the measurement of familiarity task, and the knowledge task. Then, the model selection analyses were conducted. Specifically, based on a maximum likelihood approach, I first investigated which one out of three inference models (familiarity-matching, familiarity heuristic, or knowledge-based inference) each participant would select. Next, based on Bayesian model weight, I clarified the strength of evidence for the model selection. According to the framework of attribute substitution, I classified the questions for a relationships-comparison task into "difficult" and "easy" questions when conducting data analyses.

3.1 Method

3.1.1 Participants

Ninety Japanese undergraduate students (26 participants from Aoyama Gakuin University; 64 participants from Chiba University) participated in this study; $M_{\text{age}} = 21.14$, $SD_{\text{age}} = 8.75$, and 52 participants were female. All participants completed the following tasks in about 70 minutes.

3.1.2 Tasks and materials

The three tasks were conducted — *the binary choice task* of the relationships-comparison task, *the measurement of familiarity task*, and *the knowledge task* — using a computer.

The binary choice task. Participants answered 120 binary choice questions in the format of a relationships-comparison task, such as "Which country is Sikasso in, Mali or Switzerland?" For each question, participants also rated the perceived level of difficulty, which indicated how difficult participants felt to make a correct inference. As materials for the questions, 20 countries and 120 cities were selected (see Supplementary material 1 for the procedure used to generate the questions).

The measurement of familiarity task. Participants rated how familiar they were with each object presented in the binary choice task (20 countries and 120 cities).

The knowledge task. Participants answered multiple choice questions about each object presented in the binary choice task (20 countries and 120 cities). This task consisted of four questions: a "*country*" question, such as "Which country do you think that the city is in?"; a "*region*" question, such as "Which region do you think that the city/country is in?"; a "*language*" question, such as "Which language do you think is mainly spoken in the city/country?"; and a "*religion*" question, such as "Which religion do you think is mainly followed in the city/country?"

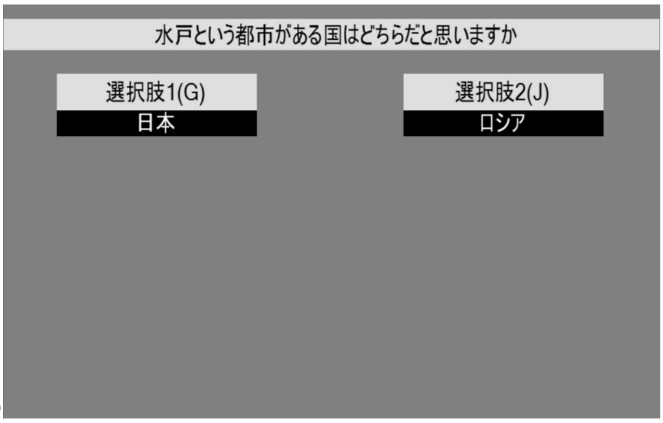
3.1.3 Procedure

In the binary choice task, at first, fixation points (two asterisks) were presented for one second at the places where country names (two alternatives) would be presented later. Then, a question and two alternatives were presented on a computer display. Participants were asked to choose one of the two alternatives by pressing one of two keys on the keyboard assigned to the alternatives ("G" key or "J" key; Fig. 7, (A)). Participants' keypress responses and their response times (i.e., the time from appearance of the question to her/his keypress) were recorded. After choosing one alternative, participants were asked to rate the difficulty level of the question using a visual analog scale. Their responses were recorded over a range of 101 points: From 0 = "very easy" on the left end to 100 = "very difficult" on the right end. (Fig. 7, (B)). After rating difficulty, participants could go on to the next question by pressing the key assigned to the "next" button ("H" key). The order of the total of 120 questions was randomized for each participant. After participants finished answering 60 binary choices and difficulty ratings, a break time was inserted. Participants could start the rest of the binary choice task at their pace by clicking the "next" button. At the time of starting or restarting the task (i.e., just before the 1st and 61st question, respectively), four filler binary choice questions and difficulty ratings were inserted as exercise trials.

In the measurement of familiarity task, a city name or a country name was presented individually on a computer display. Participants rated its familiarity level using a visual analog scale, and their responses were recorded over a range of 101 points: From 0 = "do not know at all" on the left end of the scale to 100 = "know much" on the right end of the scale.[vii] (Fig. 7, (C)). After rating the familiarity of one object, participants could go on to the next rating by clicking the "next" button on the display. When participants finished rating 70 objects, a break time was inserted. Participants could start the rest of the measurement of familiarity at their pace by clicking the "next" button. As in the binary choice task, the measurement of familiarity with five objects was inserted as an exercise trial when participants started or restarted the task. The order of the total of 140 questions (for 20 countries and 120 cities) was randomized for each participant.

In the knowledge task, a city name or a country name was presented individually on a computer display. When a city name was presented, participants answered the four questions (Fig. 7, (D)). When a country name was presented, participants answered the three questions other than the "country" question (Fig. **7**, (E)). For each question, participants were asked to choose an alternative that they thought was the most plausible. When participants had no knowledge, they could choose "I do not know." After answering these three or four questions, they could go on to the next question by clicking the "next" button on the display. When participants ended the 50^{th} and 100^{th} trials, the break times were inserted. Participants could start the next trial at their pace by clicking the "next" button. Four trials (three city names and one country name were presented) were inserted as exercise trials when participants started or restarted the task. The order of the total of 140 questions (for 20 countries and 120 cities) was randomized for each participant.

Note that, the reason why I used a visual analog scale in the task (i.e., for measuring subjective difficulty and familiarity) was to enable participants to rate difficulty or familiarity intuitively. It was expected that people generally judged which strategies they would use just in a few seconds in a task. In fact, post-hoc analyses revealed that participants made an inference within at most about 4.50 seconds even at the beginning of the difficult question. If I used another way of rating such as n-points Likert scale, then it would be difficult for participants to rate difficulty and familiarity intuitively. Based on such considerations, I adopted a visual analog scale in the current experiment.

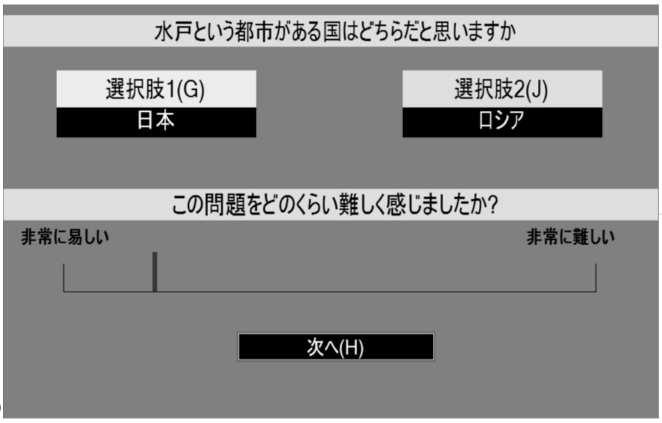


(A)

Which country is Mito city in?

Alternative 1 (G) Japan

Alternative 2 (J) Russia

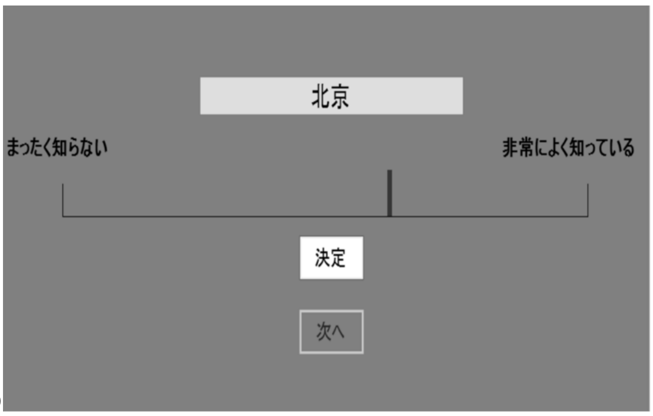


(B)

How difficult was this question?

very easy | very difficult

Next (H)



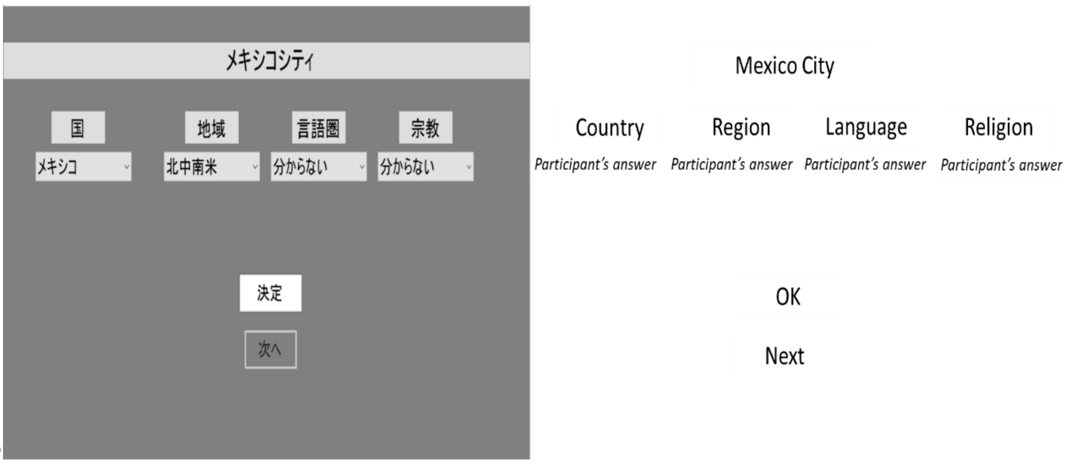
(C)

Beijing

Do not know at all | know much

OK

Next

(D) 

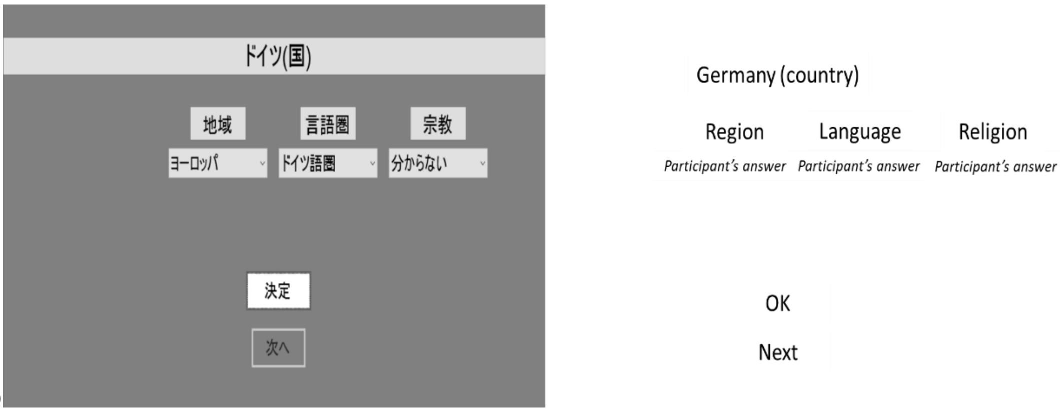
(E) 

Fig. 7 Sample pictures of the displays in Study 1. Captions on the right side of the pictures indicate English transcriptions for each description. (A) The binary choice task (an alphabet in parentheses denotes the key assigned to the alternative or button). (B) Difficulty rating task. (C) The measurement of familiarity. (D) and (E) The knowledge task (pictures show the questions of city version and country version, respectively).

Note: All sample pictures are taken from exercise trials.

3.2 Results & Discussion

sAccording to the attribute substitution framework (see the beginning of this chapter), the "difficult" and "easy" questions were defined with the following procedure. Difficulty ratings of 120 questions were transformed into z-scores for each participant. Then, the means of z-scored difficulty ratings were calculated for each question. The median of these 120 mean ratings was also computed. Finally, the higher 60 questions in its z-scored difficulty were defined as "difficult" questions, and the lower 60 questions as "easy" questions. Note that, hereafter I will report descriptive statistics related to main results. For experimental data of Study 1, see Supplementary material 2.

First, for the manipulation check, I focused on the distribution of the difficulty ratings for 120 questions. As a result, although the distribution was not a normal distribution (W = 0.89, p < .01, Shapiro-Wilk test), it was not extremely skewed or bimodal (Fig. 8; median = 0.16, 1st quantile = – 0.41, 3rd quantile = 0.60). This result indicates that some questions were difficult to answer and others were easy to answer in the relationships-comparison task, and thus my materials for this binary choice task would be appropriate in terms of difficulty.

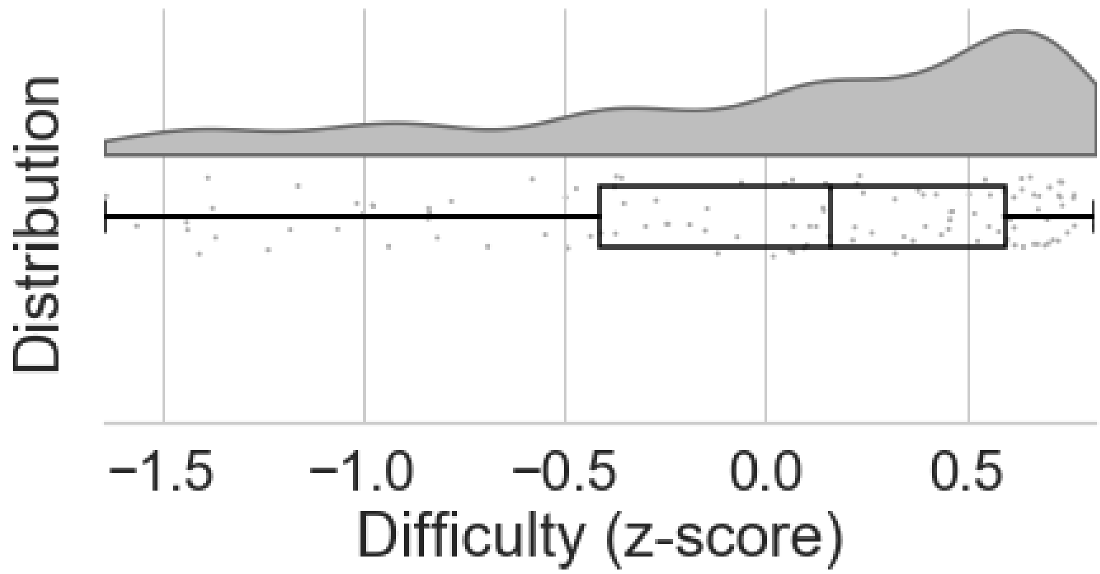


Fig. 8 The raincloud plot of the difficulty ratings (z-scored for each participant) in the binary choice task, for the manipulation check in Study 1.

3.2.1 Strategy classification by a maximum likelihood approach

For each participant, I classified the participant's inferential strategy in the binary choice task. The best model that could explain her/his inferential patterns was identified using a maximum likelihood approach (e.g., Hilbig, Michalkiewicz, Castela, Pohl, & Erdfelder, 2015; Pachur & Aebi-Forrer, 2013; Pachur, Hertwig, & Rieskamp, 2013). I calculated the goodness of fit, G^2, of a strategy k (FM, FH, or KI) for participant i using the following equation:

$$G_{i,k}^2 = -2\sum_{x=1}^{N} \ln[f_x(y)] \tag{1}$$

where $f_x(y)$ denotes the probability that a strategy predicts her/his inference y in a binary choice question x. If the strategy k required guessing, then $f_x(y) = 0.5$; otherwise, if an observed inference in the question accorded with the strategy's prediction, then $f_x(y) = 1 - \varepsilon_{i,k}$; and if it did not accord, then $f_x(y) = \varepsilon_{i,k}$. Here $\varepsilon_{i,k}$ denotes participant i's application error across all N questions (in Study 1, $N = 60$ in both difficult and easy questions) for strategy k. $\varepsilon_{i,k}$ was estimated as the proportion of her/his inferences that deviated from strategy k's prediction, which represented the maximum likelihood estimate of this parameter (Bröder & Schiffer, 2003). In this measure, a lower G^2 indicates the better model fit[viii].

For each difficulty level and each participant, the best inferential model was identified with the following procedure. For the heuristic models (FM and FH), I first estimated the best decision threshold based on G^2. The decision threshold was applied to the absolute value of familiarity with objects:"$||\mathrm{FamQ} - \mathrm{FamA}| - |\mathrm{FamQ} - \mathrm{FamB}||$" for FM; and "$|\mathrm{FamQ} - \mathrm{medianFamQs}|$ and $|\mathrm{FamA} - \mathrm{FamB}|$" for FH. For each participant, G^2 was calculated for 100 patterns of threshold values by a grid search (i.e., from 1 to 100). I then defined the threshold value showing the lowest G^2 as her/his best threshold. On the other hand, for the knowledge model (KI), I first identified her/his best LEX (i.e., the lowest G^2 among 3! = 6 patterns) and calculated G^2 of TAL 1 and that of TAL 2. I then defined a model whose G^2 was the lowest among the best LEX, TAL 1, and TAL 2 as her/his best KI.

In the following analyses, I used data of the best FM, FH, and KI for each participant. I show the mean and SD of the decision thresholds in FM or FH[ix], the classified rates (based on Bayesian model weight; described later) (Table 1, (A)), and the number of participants who were classified into LEX, TAL, or Both (i.e., the same G^2 value) for each difficulty level (Table 1, (B)). The total classified rates were .90 (81/90) for difficult questions and .99 (89/90) for easy questions, respectively. Note that, six LEX models and two TAL models in this thesis had very similar algorithms to each other, so it was not so strange that some participants were classified as “Both.” In fact, the mean similarity rates of prediction in 12 patterns (6 LEX patterns * 2 TAL patterns) were .96 in difficult questions and .95 in easy questions.

Next, I examined which model could best explain inferences among FM, FH, and KI for each participant. In particular, I compared fitness among these three models and selected one model that explained her/his inference patterns best. Here, as an index of strength of evidence for model selection, the Bayesian model weight, w_M (e.g., Hilbig et al., 2015; Honda et al., 2017; Jenny et al., 2014) was adopted, based on individual Bayesian information criteria (BIC). w_M was calculated for each participant. By this analysis, I could rigorously examine the superiority of one model over the other models (Jenny et al., 2014). w_M was calculated as follows:

$$w_M = \frac{\exp\left(-\frac{1}{2}\Delta BIC_M\right)}{\sum_i \exp\left(-\frac{1}{2}\Delta BIC_i\right)} \tag{2}$$

$$\mathrm{BIC} = G^2 + p * \ln(n) \tag{3}$$

As for w_M, ΔBIC_M denotes the difference between model M and the best model; and ΔBIC_i denotes the BIC difference between the best model and one specific model i in the comparison. As for BIC, p and n denote the number of free parameters and the number of choice pairs, respectively. Especially in this study, because the model weight of the best model was focused on, model M was always regarded as the best model. Therefore ΔBIC_M = |the best model BIC – the best model BIC|, and $\exp(-\frac{1}{2}\Delta BIC_M) = 1$. In addition, ΔBIC_i = |the best model BIC – FM BIC|, |the best model BIC – FH BIC|, or |the best model BIC – KI BIC|. That is, $\sum_i \exp(-\frac{1}{2}\Delta BIC_i)$ was calculated for normalization. Furthermore, only one parameter, ε, for each strategy in calculating G^2, so $p = 1$ (fixed); and the number of questions was 60 both for difficult and for easy, so $n = 60$ (fixed).

According to previous studies (e.g., Honda et al., 2017; Raftery, 1995), the evidence for classification was assumed in the following way: $w_M \geq .99$ as “very strong” evidence, $.95 \leq w_M < .99$ as “strong” evidence, $.75 \leq w_M < .95$ as “positive” evidence, and $.50 \leq w_M < .75$ as “weak” evidence. If $w_M < .50$, or the values of w_M were equal between two or more models, one model could not explain her/his inferential patterns. If so, then s/he was defined as “Not classified.”

3.2.2 Main results regarding strategy classification

The results of the Bayesian model weight approach for difficult and easy questions are shown in Table 1 (C) and (D), and also Fig. 9. In the following analyses of Study 1, the data of participants who were classified into one of the three models were used. Hereafter, participants who were classified into FM, FH, or KI were defined as "FM users," "FH users," or "KI users," respectively, and also those into FM or FH as "heuristic users" and those into KI as "knowledge users."

In terms of participants' inferential strategies, FM was selected by as many participant as FH in difficult questions (32 participants). However, FM was selected with stronger evidence than FH. The classified rates of "very strong" evidence (i.e., $.99 \leq w_M$) and "strong" evidence (i.e., $.95 < w_M \leq .99$, respectively) were respectively .13 and .19 in FM, whereas .03 and .16 in FH. These result indicate that FM was more typically used in a relationships-comparison task, that is, people tended to make inferences based not only on familiarity between two alternatives (i.e., FH) but also on a similarity in familiarity among objects (i.e., FM), especially in difficult questions. It is suggested that familiarity with objects could be used as an inference cue for the relationships-comparison task when people did not have sufficient knowledge.

Table 1 Results in Study 1. (A) The mean and SD of the best decision threshold, and the classified rates for FM and FH. (B) The classified rates for KI. (C) Classified rates of the evidence of classification based on Bayesian model weight in difficult questions. (D) Those in easy questions.

Note: FM: familiarity-matching model. FH: familiarity heuristic model. KI: knowledge-based inference model. LEX: lexicographic model. TAL 1: Tally 1 model. TAL 2: Tally 2 model. w_M: Bayesian model weight.

Note for (A) and (B): Values in parentheses denote "the number of participants classified into the model divided by the number of classified participants." The word "Both" in (B) means that values of $\boldsymbol{G^2}$ in the best LEX model were the same as those in the best TAL model.

Note for (C) and (D): Values in parentheses denote "the number of classified participants divided by the number of total participants."

(A) Mean and SD of decision thresholds (classified participants)

Difficult questions				Easy questions			
FM; .40 (32/81)		FH; .40 (32/81)		FM; 0 (0/89)		FH; .08 (7/89)	
Mean	SD	Mean	SD	Mean	SD	Mean	SD
16.5	15.1	3.72	3.32	---	---	7.57	5.98

(B) The numbers of the selected models in KI (classified participants)

Difficult questions			Easy questions		
KI; .21 (17/81)			KI; .92 (82/89)		
LEX	TAL	Both	LEX	TAL	Both
.65 (11/17)	.12 (2/17)	.24 (4/17)	.57 (47/82)	.09 (7/82)	.34 (28/82)

(C) Strategy classification and its evidence in difficult questions

Not Classified (N.C.) rate	.10 (9/90)		
	FM	FH	KI
Classified rate	.40 (32/81)	.40 (32/81)	.21 (17/81)
Very strong [.99 ≤ w_M]	.13 (4/32)	.03 (1/32)	.29 (5/17)
Strong [.95 < w_M ≤ .99]	.19 (6/32)	.16 (5/32)	.12 (2/17)
Positive [.75 < w_M ≤ .95]	.31 (10/32)	.47 (15/32)	.35 (6/17)
Weak [.50 < w_M ≤ .75]	.38 (12/32)	.34 (11/32)	.24 (4/17)

(D) Strategy classification and its evidence in easy questions

Not Classified (N.C.) rate	.01 (1/90)		
	FM	FH	KI
Classified rate	0 (0/89)	.08 (7/89)	.98 (82/89)
Very strong [.99 ≤ w_M]	0	0 (0/7)	.80 (66/86)
Strong [.95 < w_M ≤ .99]	0	0 (0/7)	.05 (4/86)
Positive [.75 < w_M ≤ .95]	0	.57 (4/7)	.10 (8/86)
Weak [.50 < w_M ≤ .75]	0	.43 (3/7)	.05 (4/86)

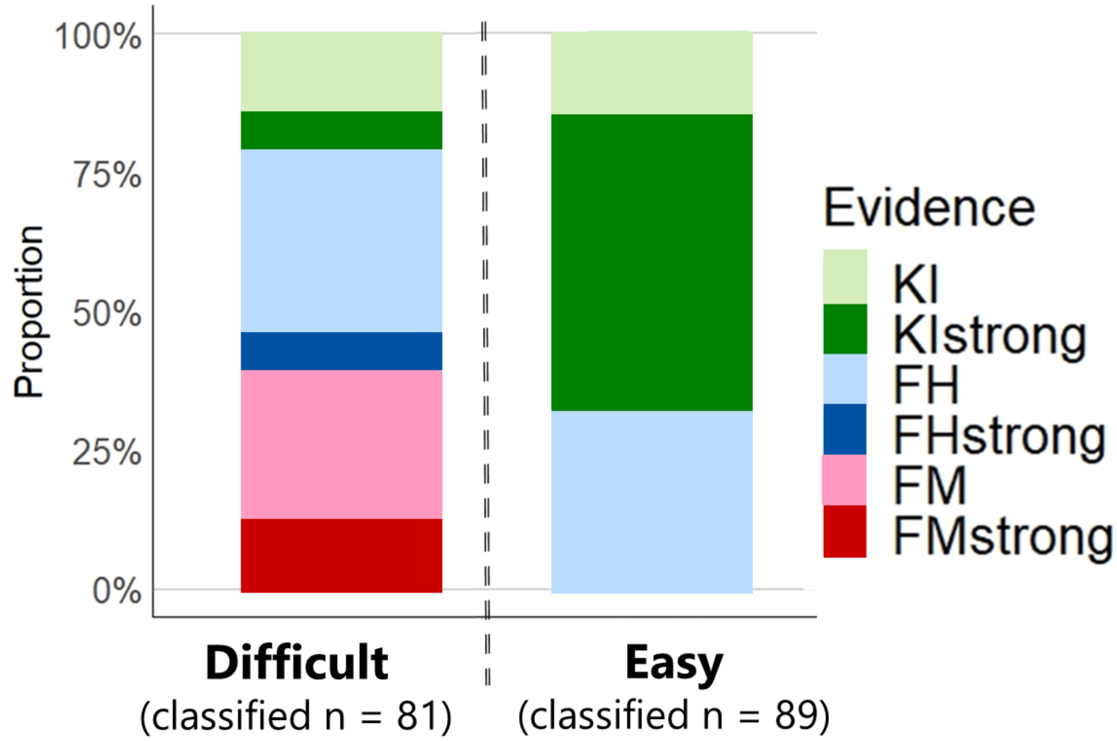


Fig. 9 Proportions of the strength of evidence for model selection in Study 1 (visualizing Table 1 (C) and (D)).

Note: FM: familiarity-matching model. FH: familiarity heuristic model. KI: knowledge-based inference model. The word “strong”: “very strong” evidence (i.e., .99 ≤ w_M) and “strong” evidence (i.e., .95 < w_M ≤ .99).

Additionally, I also focused on the shifts of participants' heuristics-based or knowledge-based strategies depending on the difficulty levels, in terms of attribute substitution (see the beginning of this chapter; see also Honda et al., 2017). In difficult questions, 81 out of 90 participants were classified into one of the three models, and heuristic models (i.e., FM and FH) were selected by significantly more participants than the knowledge model was (heuristic (32+32)/81 = .79; knowledge 17/81 = .21; $p < .001$, 95%CI = [.69, .87], binomial test). In easy questions, on the other hand, 89 out of 90 participants were classified into one of the three models, and the knowledge model was selected more often than the heuristic models were (heuristic (0+7)/89 = .08; knowledge 82/89 = .92; $p < .001$, 95%CI = [.03, .15], binomial test). To confirm the tendencies in individual levels, I conducted additional analyses. I found that 80 participants were classified into one of the three models in both difficulty levels, and that 56 out of these 80 participants were classified as heuristic users in difficult questions and as knowledge users in easy questions (Table 2). These results show that participants were likely to shift their strategies depending on their subjective difficulty of the questions, which was consistent with the prediction from the attribute substitution (e.g., Honda et al., 2017).

Table 2 The shift of inferential strategies (heuristics [i.e., FM or FH] or knowledge [i.e., KI]) depending on the difficulty level, for each participant (n = 80, who were classified into one of the three models both in difficult questions and in easy questions). The values denote "the number of participants divided by 80".

	Easy -- Heuristic	Easy -- Knowledge
Difficult – Heuristic	7 / 80	56 / 80
Difficult -- Knowledge	0 / 80	17 / 80

Note that, according to many previous studies (e.g., Hilbig & Pohl, 2009; Pachur & Hertwig, 2006), I also analyzed participants' response times for making inferences using a mixed linear model. I did not have specific hypotheses about response time, therefore I conducted the analyses explanatorily, focusing on the difference between heuristic and knowledge users. As a result, the order effect (i.e., the tendency that response time became shorter as the questions proceeded; e.g., Schweickart & Brown, 2014) was observed in difficult questions. However, there were no significant differences between heuristic and knowledge users. For the results of the analyses of response time in detail, see Supplementary material 3.

3.2.3 Similarity between inference models

Even if each model is constructed based on operationally different assumptions, predictions by these models for the real-world objects are sometimes highly similar to each other. Previous studies (e.g., Hertwig, Pachur, & Kurzenhäuser, 2005; Pachur, Hertwig, Gigerenzer, & Brandstätter, 2013) reported that some inferential models they had proposed showed very similar predictions about social statistics, although each model was assumed to describe differential strategies. Then, I calculated the similarities of prediction between models FM and FH, FM and KI, and FH and KI for each participant. Fig. 10 shows the proportion of accordance rates for the prediction between models. In both difficulty levels, the mean accordance rates of predictions by two heuristic models were not so similar to each other (FM vs. FH: .59 in difficult questions; and .57 in easy questions). Based on the results, although FM and FH seemed to be very similar to each other in terms of their algorithms, FM could be regarded as an essentially different model from FH. The predictions of the knowledge model were also not similar to those of these heuristic models. Especially in easy questions, these mean accordance rates were highly dissimilar (FM vs. KI: .36; and FH vs. KI: .36 respectively). In sum, these results indicate that predictions of inferential patterns by the three models were not so similar to each other.

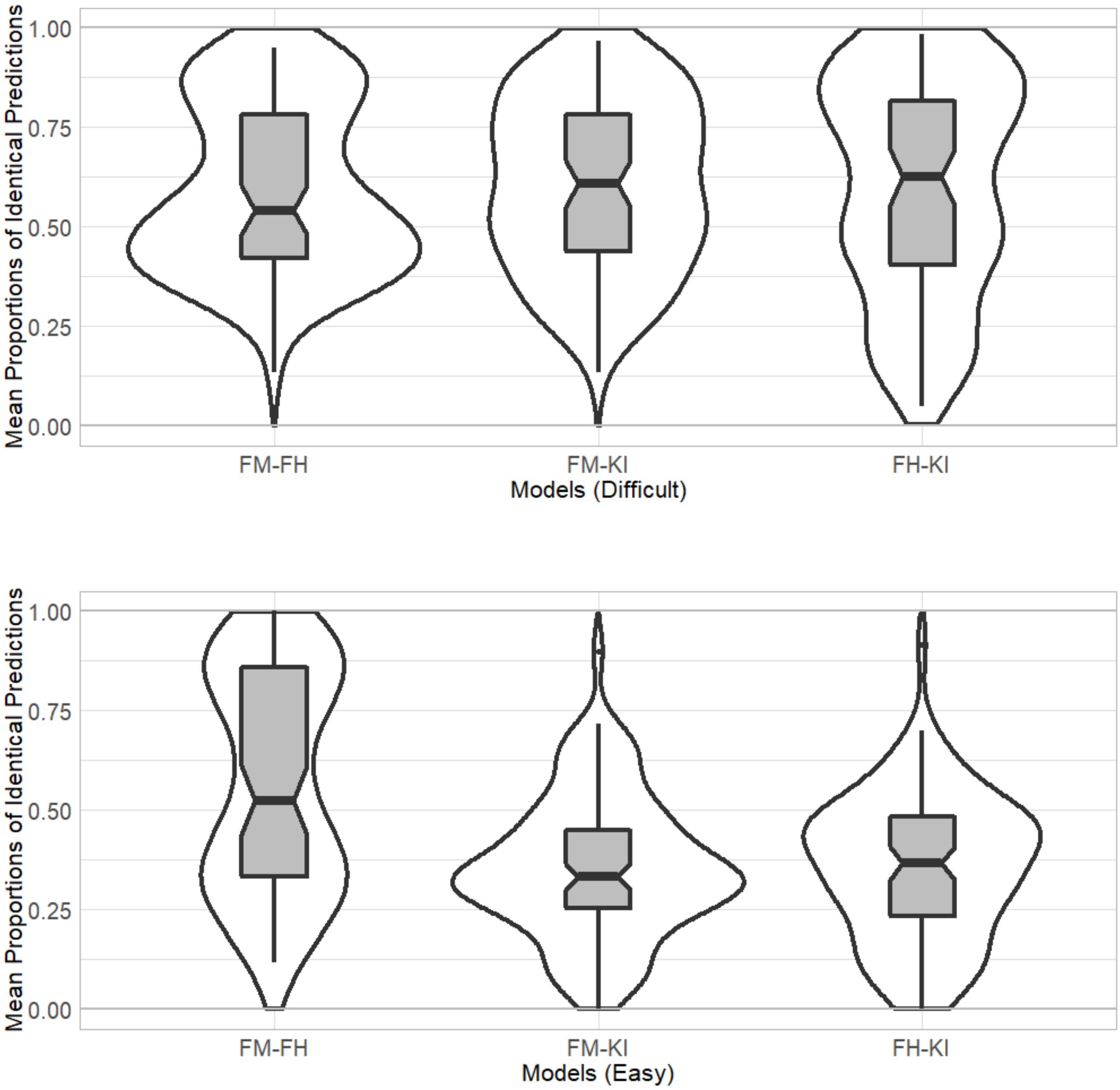


Fig. 10 Similarity between inferential models. The mean proportions of the same predictions between models in difficult questions (upper panel) and easy questions (lower panel). Each graph shows the distribution of the mean proportion of the same predictions between two models, FM and FH (left), FM and KI (middle), and FH and KI (right), respectively.

3.3 Summary of Study 1

The human inferential strategy that would be used in a new task structure of binary choice (i.e., *relationships-comparison task*) was investigated by constructing three inferential models — *familiarity-matching* (FM), *familiarity heuristic* (FH), and *knowledge-based inference* (KI). I found that participants tended to use heuristics, especially a new type of heuristic, FM, with strong evidence in difficult questions. Moreover, participants were likely to shift their inferential strategies depending on the difficulty level of the questions, which was consistent with the prediction from the attribute substitution. Based on the results in Study 1, I could confirm that FM was typically used in a relationships-comparison task.

Chapter 4 Study 2: Examinations of ecological rationality (Behavioral experiment, data analyses, and computer simulation)

Study 1 showed that people tended to use a simple heuristic, especially familiarity-matching (FM), for a relationships-comparison task. As for heuristic models, based on the results of model selection analyses in Study 1, FM would be used with the stronger evidence than FH. However, it remains unclear whether FM is an ecologically rational and accurate strategy[x]. That is, I still do not know how well the structure of FM will fit to that of the real-world environment, and whether using FM will lead correct (accurate) inferences. In Study 1, I used one question set picked up by an experimenter, and therefore the experimental materials might not be appropriate for investigating the real-world environment. Thus, in Study 2, I should use new materials without experimenter's own criteria for picking up them.

As described in Chapter 1, many memory-based heuristics can have *ecological rationality* because subjective memory experiences (e.g., recognition, fluency, or familiarity) are often positively correlated with a particular criterion for making inferences (e.g., population size). In a population inference task, for example, larger cities are frequently appeared in media (e.g., newspapers). Then, people are likely to get more familiar with larger cities since they often see or hear the names in the real world. Because of such structures of an environment, memory-based heuristics are likely to lead correct inferences (e.g., Goldstein & Gigerenzer, 2002; Honda et al., 2017; Schooler & Hertwig, 2005).

Based on these previous studies, I expected the following three relationships in the real-world structure for a relationships-comparison task. First, the frequency of appearance of objects (i.e., city names and country names) is highly correlated with people's familiarity with objects (a right arrow in Fig. **1** (D)). As in Chapter 1, consider the question: "Which country is Valencia in, Spain or Bolivia?" If a city name "Valencia" is often mentioned in media, then people will get more familiar with Valencia because they often see or hear the name (i.e., as in Fig. **1** (B) right arrow). Second, the frequency of appearance of cities is highly correlated with that of the countries where the cities are located (two left arrows in Fig. **1** (D)). Typically, it will be assumed that a familiar city tends to be in a familiar country, and vice versa. For example, generally, a city name "Valencia" may be frequently mentioned in many media articles (e.g., a name "Valencia orange"); and a country name "Spain," where Valencia is located, may also be frequently mentioned in many articles (e.g., international news, sight-seeing guides, etc.). In contrast, a city name "El Alto" may be less likely to be mentioned (and also a country name "Bolivia" may be less frequently mentioned). Third, familiar cities tend to be judged in familiar countries (a lower arrow in Fig. **1** (D)). For example, because a city name "Valencia" as well as a country name "Spain" are frequently mentioned in media (i.e., second relationship in the above) and people are more likely to be familiar both with Valencia and Spain (i.e., first relationship in the above), then considering that Valencia is in Spain based on a similarity in familiarity can be a correct inference (i.e., FM will be an ecologically rational strategy).

In Study 2, in terms of the extent to which a heuristic reflects environmental structures (i.e., *ecological rationality*; e.g., Brighton, 2020; Gigerenzer, 2008; Gigerenzer & Gaissmaier, 2011; Goldstein & Gigerenzer, 2002; Spiliopoulos & Hertwig, 2019), the accuracy of the heuristic in a relationships-comparison task was examined through a behavioral experiment and a computer simulation, by using new experimental materials "the 50 countries with the highest populations in the world and their 50 capitals." It is considered that the materials in Study 2 had less arbitrary aspect and reflected the real-world environmental structure more appropriately than those in Study 1, because there were no experimenter's own criteria in picking up cities and countries.

The main procedure of Study 2 was as follows. In a behavioral experiment, people's familiarity of each object (i.e., 50 cities and 50 countries) was measured. Next, by using online databases, I investigated how often people would see or hear these city names or country names in the real world. Specifically, I used the hit number of searching as an index for the frequency of seeing or hearing the names. Then, I examined whether frequently appeared objects in the real world (e.g., objects with high hit numbers) were generally familiar to people, and whether more frequently appeared cities would be in more frequently appeared countries. In a computer simulation, finally, I generated hypothetical relationships-comparison tasks (i.e., all possible pairs using the materials), and then calculated the accuracy (i.e., correct rate) for the hypothetical questions assuming that participants constantly used FM in all questions.

4.1 Participant, material & procedure (behavioral experiment)

In a behavioral experiment in Study 2, thirty-nine Japanese undergraduate students (from Chiba University) participated; M_{age} = 18.46, SD_{age} = 0.80, and 25 participants were female. No participants had participated in the experiment of Study 1.

As an experimental material in Study 2, the 50 countries with the highest populations in the world and their 50 capitals were used. Participants were asked to rate the familiarity of each of the 100 objects (i.e., 50 cities and 50 countries) using a visual analog scale, in the same way as in the measurement of familiarity task in Study 1.

4.2 An environmental structure in the real world (data analyses)

I analyzed the correlation between the frequency of appearance of objects (i.e., city names and country names) and people's familiarity with them, and that between the frequency of appearance of city names and that of country names. As an index for the object's frequency of appearance in the real-world environment, I used the mean number of hits in two online databases of Japanese newspapers: *Kikuzo II Visual* (an online database of Asahi Shimbun; date range: January 1, 1984 to May 23, 2016) and *Yomidasu Rekishikan* (an online database of Yomiuri Shimbun; date range: January 1, 1986 to May 23, 2016). When I searched for objects (50 cities and 50 countries) in both databases, I traced from the oldest to the newest date in the national news.

4.3 Calculation of accuracy of familiarity-matching (computer simulation)

The accuracy (i.e., correct rate) of FM in the hypothetical binary choices was analyzed, using familiarity ratings for the 50 countries and 50 cities collected from 39 participants in the behavioral experiment. The correct rate of FM was calculated with the following procedure.

1. Hypothetical relationships-comparison tasks such as "Which country is city Q in, country A or country B?" were generated, and it was assumed that each problem "was inferred" based on individuals' FamQ, FamA and FamB.
2. The algorithm of FM was applied for each question (see section 2.1). Note that a decision threshold was not considered in Study 2. If differences of familiarity were equal to each

other (i.e., |FamQ − FamA| = |FamQ − FamB|), then it was assumed that one of two alternatives was randomly chosen.

3. The above two steps were applied to all possible 50 * 49 questions (i.e., 50 pairs of "cities and correct countries" * the remaining 49 false countries) using the familiarity ratings provided by the participants in the behavioral experiment, and then calculated individuals' mean of the correct rate.

4.4 Results & Discussion

As the data processing for the hit numbers of objects (i.e., an index for the object's frequency of appearance in the real world), I transformed the numbers of hits into log-scales[xi] and then converted them into z-scores. As to the familiarity with objects, I used the mean of participants' familiarity ($N = 39$) with each object, and then converted them into z-scores to make these familiarity ratings the same scale as the hit numbers.

In the data analyses, I first calculated the correlation between the frequency of appearance of objects (i.e., city names and country names) and people's familiarity with them, in order to confirm a previously reported positive correlation between people's media exposure and familiarity (e.g., Goldstein & Gigerenzer, 2002). Pearson's correlation coefficient was $r = .88$ ($p < .001$; 95%CI = [.82, .92]; Fig. 11, upper right). Therefore, I could confirm the tendency that the more frequently an object appeared in the media, the more familiar with the object people would be. This result was consistent with previous studies (e.g., Goldstein & Gigerenzer, 2002; Schooler & Hertwig, 2005).

Next, I also calculated the correlation between the frequency of appearance of city names and that of country names. Pearson's correlation coefficient was $r = .86$ ($p < .001$; 95%CI = [.77, .92]; Fig. 11, upper left). Therefore, the frequency of the appearance of cities in media was highly correlated with that of the corresponding countries.

Finally, the accuracy (i.e., correct rate) of FM by a computer simulation was calculated. Lower panels in Fig. 11 show the distributions of 39 participants' accuracy rates for FM, in a form of (left) individual data and of (right) empirical cumulative distribution function. In individual data, the x-axis denotes participants (ordered by their correct rates), and the y-axis denotes their mean of accuracy rates. In empirical cumulative distribution data, the x-axis denotes the accuracy rate, calculated by FM and the y-axis denotes the frequency. The dotted vertical lines indicate the chance levels (.50). The solid vertical lines indicate the mean accuracy rates (.67). In FM, the accuracy rates exceeded the chance level for all participants. The mean of their accuracy rates was significantly higher than the chance level (mean: .67; $V = 780$, $p < .001$, $r = 0.79$, One-sample Wilcoxon rank sum test).

Taken together, the results obtained here show that the new type of heuristic, FM, will be an ecologically rational heuristic for a task structure of a relationships-comparison task. As for FM, people consider the relationships between familiarity of a city and that of a country as an inference cue. This structure of inference can effectively reflect the real-world environmental structure, and therefore FM will be likely to make correct inferences (see Fig. 11).

Note that, one may doubt the appropriateness of using the hit number in online newspaper databases as the index of the frequency of appearance (i.e., the frequency that people see or hear the names) in the real world. As described Chapter 1, I investigated the adaptive use of heuristics (especially in Study 2, the *accuracy* of familiarity-matching) according to the adaptive toolbox framework proposed by Gigerenzer and his colleagues (e.g., Gigerenzer et al., 1999). Their previous works mainly focused on the specific domain such as general knowledge that is seen or heard through media (e.g., city population). So, I also followed their framework and their ways of analyses. In this study, it was expected that general knowledge about city names or country names could generally be seen through media and documents in the real world. Therefore, I believe that it would be appropriate to use the hit number of databases.

4.5 Summary of Study 2

Study 2 focused on the accuracy of a heuristic in terms of *ecological rationality* in a relationships-comparison task, using the materials that were different from Study 1's materials. Through a behavioral experiment and a computer simulation, I clarified that familiarity-matching would be an ecologically rational strategy because it could well reflect the real-world environmental structure. Specifically, the more frequently an object appears, the more familiar with the object people are. Additionally, the more frequently a city name appears, the more frequently the corresponding country name tends to appear in the real-world environment. Thus, a more familiar city is often in a more familiar country and, therefore, FM can have ecological rationality in a relationships-comparison task.

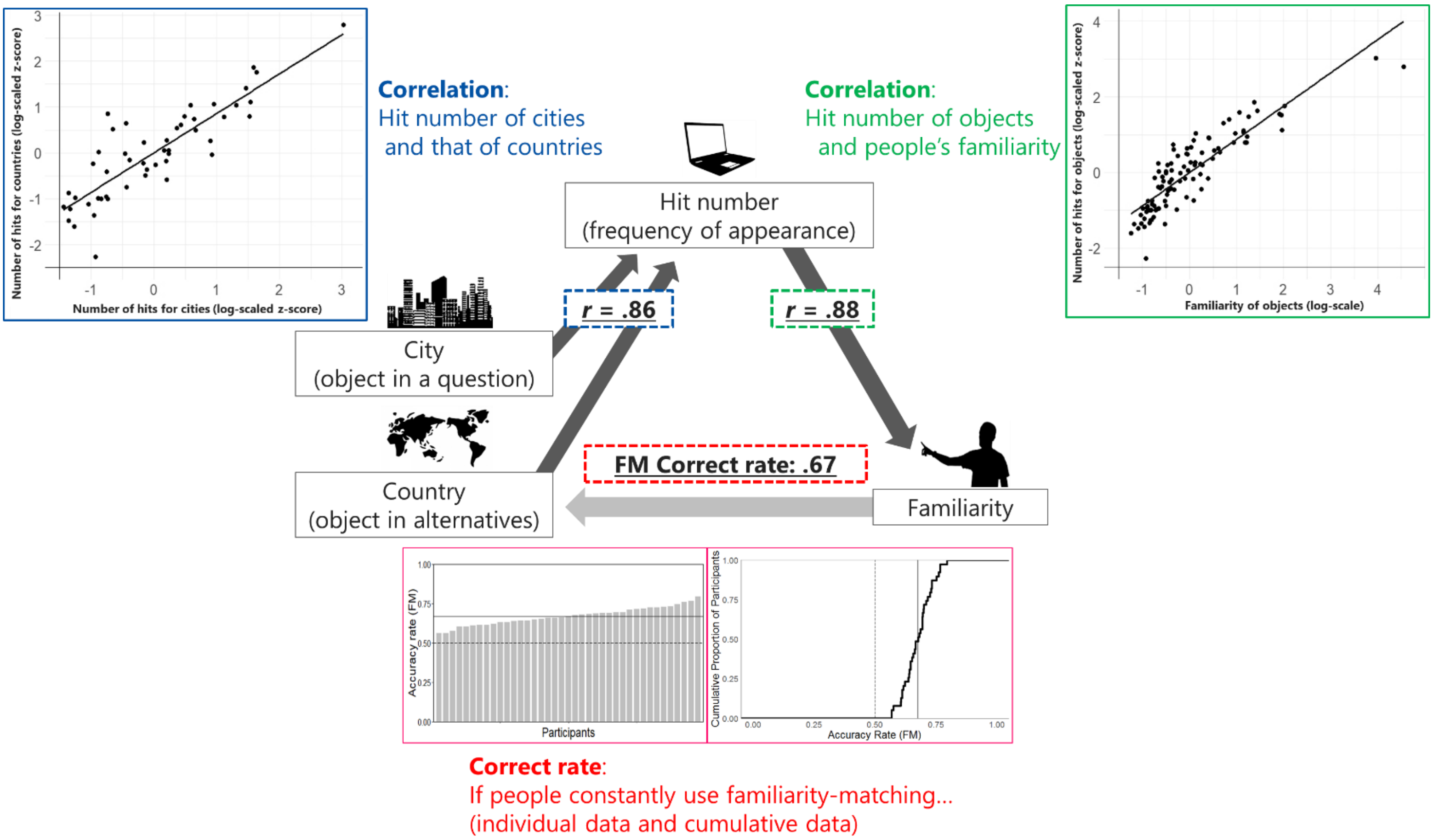


Fig. 11 An environmental structure that people may exploit and Results in Study 2. (Upper right; green font) Hit number of objects (log-scaled z-score) is highly correlated with people's familiarity (z-score). (Upper left; blue font) Hit number of cities is highly correlated with that of the corresponding countries. (Lower; red font) Each participant's mean accuracy rates calculated by a computer simulation are shown in a form of (left) individual data and of (right) empirical cumulative distribution function. The dotted vertical lines indicate the chance levels (.50). The solid vertical lines indicate the mean accuracy rates (.67). In this simulation, we calculated each participant's accuracy rate assuming that they constantly used familiarity-matching based on the familiarity ratings provided from the behavioral experiment.

Chapter 5 Study 3: Replication of Study 1 (Behavioral experiment)

Some researchers may argue that the ecological rationality of familiarity-matching (FM), shown in Study 2, does not necessarily support Study 1's findings that people often used FM in difficult relationships-comparison tasks. This is because the experimental materials I used were different between Study 1 and Study 2: In Study 1, objects (city names and country names) were picked up by an experimenter based on experimenter's own criteria (see Supplementary material 1), while in Study 2, objects were picked up without experimenter's criteria (i.e., based on only a rank with the highest population). To address this concern, in Study 3, I should confirm whether the main results of Study 1 could be replicated, using the materials used in Study 2.

The main procedure of Study 3 is as follows. The same materials as in Study 2 was used (the 50 countries with the highest populations in the world and their capitals) to generate a relationships-comparison task. In a behavioral experiment, participants were asked to answer the binary choice task, the measurement of familiarity, and the knowledge task, in the same way as in Study 1. Then, I investigated which inferential model (familiarity-matching [FM], familiarity heuristic [FH], or knowledge-based inference [KI]) each participant would select, based on the model selection analyses used in Study 1. I will now focus on the results of model selection, which were the main findings of Study 1.

5.1 Participant, material & procedure

Fifty-one Japanese university students (41 were from Chiba University and 10 were from the University of Tokyo) participated in Study 3; M_{age} = 19.5, SD_{age} = 1.54, and 30 participants were female. None of them had participated in either of Study 1 or Study 2.

Using the materials from Study 2, the 25 odd-number ranking countries (i.e., the 1st,

3^{rd}, ⋯, 49^{th} country with the highest population) and 100 cities were picked up (four cities were selected from each country[xii]). Then, in a behavioral experiment, the following three tasks were conducted — the binary choice task of the relationships-comparison task (100 questions), the measurement of familiarity task (100 cities + 25 countries = 125 questions), and the knowledge task (100 cities + 25 countries = 125 questions) — just as in Study 1 with the exceptions that participants completed these tasks in a form of online questionnaire and that the break times were inserted after 50^{th} question in the binary choice task, the 63^{rd} question in the measurement of familiarity task, and the 45^{th} and 85^{th} questions in the knowledge task. For details on the way of generating the binary choice questions, see Supplementary material 4.

5.2 Results & Discussion

In Study 3, I will focus on the results of model selection analyses, which was the main findings obtained in Study 1.

The "difficult" and "easy" questions (50 questions were assigned to each difficulty level) were first defined as in Study 1. That is, I regarded the 50 questions whose difficulty ratings (z-scored) were above or equal to the median as "difficult" and were below the median as "easy." And then, the model selection analyses were conducted to identify which strategy (FM, FH, or KI) participants used in a relationships-comparison task, the same as in Study 1.

For the manipulation check, I first focused on the distribution of the difficulty ratings for 100 questions, in the same way as in Study 1. As a result, although the distribution was not a normal distribution ($W = 0.88$, $p < .01$, Shapiro-Wilk test), it was not extremely skewed or bimodal (Fig. 12; median = − 0.89, 1^{st} quantile = 0.14, 3^{rd} quantile = 0.83). Thus, my materials of the binary choice tasks in Study 3 would be appropriate in terms of difficulty, as in Study 1.

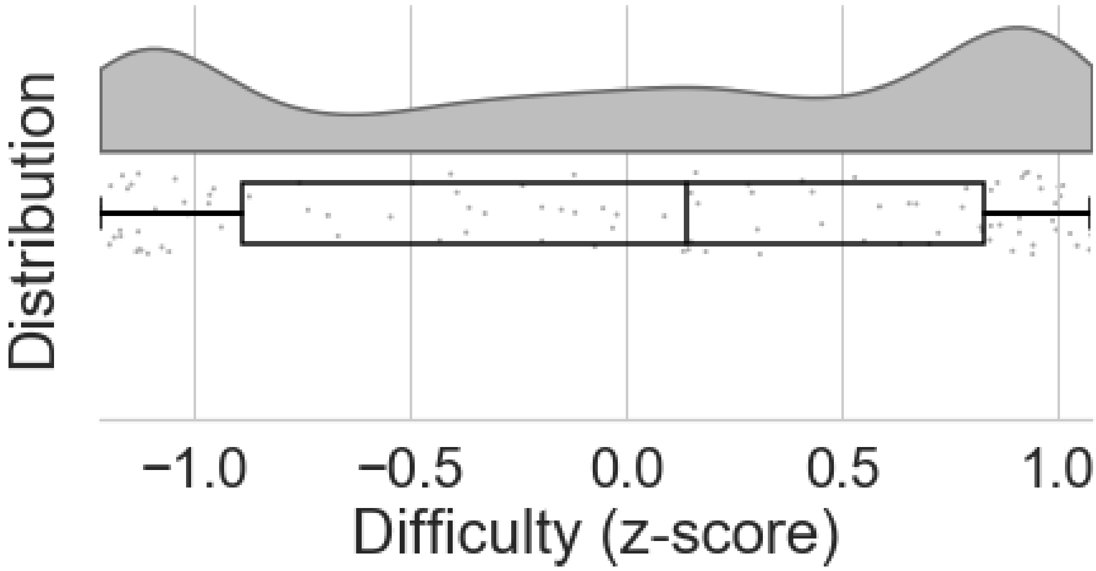


Fig. 12 The raincloud plot of the difficulty ratings (z-scored for each participant) in the binary choice task, for the manipulation check in Study 3.

Table 3 (A) and (B) show the mean and SD of the decision thresholds of FM and FH, and the classified rates and the number of participants who were classified into LEX, TAL, or Both, for each difficulty level. In difficult questions, 47 out of 51 participants were classified into one of the three models: thirty-five participants into FM, six participants into FH, and six participants into KI; and the total classified rate was .92 (47/51). The heuristic models were selected more often than the knowledge model (heuristic .87 vs. knowledge .13; $p < .001$, 95%CI = [.74, .95], binomial test). Furthermore, in the 41 heuristic users, the proportion of choosing FM was significantly higher than that of choosing FH (FM .85 vs. FH .15; $p < .001$, 95%CI = [.71, .94], binomial test). These results indicated that participants tended to make inferences based on FM in difficult questions[xiii]. In easy questions, on the other hand, all 51 participants were classified into one of the models: one participant into FM, and fifty participants into KI; and the total classified rate was 1.0 (51/51). The knowledge model was more selected than the heuristic model (heuristic .02 vs. knowledge .98; $p < .001$, 95%CI = [.00, .10], binomial test). As predicted, participants were more likely to rely on heuristics, especially FM, in difficult questions, and on their knowledge in easy questions.

Table 3 (C) and (D), and also Fig. 13 show the Bayesian model weight (w_M) calculated by individual Bayesian information criteria (BIC) for each model with the same procedure as in Study 1. In difficult questions, FM were selected by more participants with stronger evidence than FH (FM: 35 participants; the classified rates of "very strong" and "strong" are .11 and .14, respectively. FH: 6 participants; the classified rates of "very strong" and "strong" are .17 and 0, respectively). This result corroborates that FM was more robustly used than FH in a difficult relationships-comparison task. In Study 1, the number of FM users was the same as that of FH users (both 32 participants; but the classification evidence for FM was stronger than that for FH), while in Study 3, the number of FM users was much more than that of FH users (35 vs. 6 participants, respectively). Although I do not have clear evidence about this difference, it may be because the materials in Study 3 (picked up without experimenter's own criteria) reflected the real-world environmental structure more directly than those in Study 1, and therefore FM could work as a more effective strategy than FH for solving the binary choice questions.

In sum, participants' inferential patters observed in Study 3 were generally consistent with those observed in Study 1, and I can conclude that the main findings obtained in Study 1 were highly replicable.

Table 3 Results in Study 3. (A) The mean and SD of the best decision threshold, and the classified rates for FM and FH. (B) The classified rates for KI. (C) Classified rates of the evidence of classification based on Bayesian model weight in difficult questions. (D) Those in easy questions.

Note: FM: familiarity-matching model. FH: familiarity heuristic model. KI: knowledge-based inference model. LEX: lexicographic model. TAL 1: Tally 1 model. TAL 2: Tally 2 model. w_M: Bayesian model weight.

Note for (A) and (B): Values in parentheses denote "the number of participants classified into the model divided by the number of classified participants." The word "Both" in (B) means that values of $\boldsymbol{G^2}$ in the best LEX model are the same as those in the best TAL model.

Note for (C) and (D): Values in parentheses denote "the number of classified participants divided by the number of total participants."

(A) Mean and SD of decision thresholds (classified participants)

Difficult questions				Easy questions			
FM; .74 (35/47)		FH; .13 (6/47)		FM; 1.0 (1/1)		FH; 0 (0/1)	
Mean	SD	Mean	SD	Mean	SD	Mean	SD
28.3	20.4	16.7	24.3	6	0	---	---

(B) The numbers of the selected models in KI (classified participants)

Difficult questions			Easy questions		
KI; .13 (6/47)			KI; .98 (50/51)		
LEX	TAL	Both	LEX	TAL	Both
.33 (2/6)	0 (0/6)	.67 (4/6)	.42 (21/50)	.10 (5/50)	.48 (24/50)

(C) Strategy classification and its evidence in difficult questions

	FM	FH	KI
Not Classified (N.C.) rate	.08 (4/51)		
Classified rate	.74 (35/47)	.13 (6/47)	.13 (6/47)
Very strong [.99 ≤ w_M]	.11 (4/35)	.17 (1/6)	.17 (1/6)
Strong [.95 < w_M ≤ .99]	.14 (5/35)	0 (0/6)	.33 (2/6)
Positive [.75 < w_M ≤ .95]	.46 (16/35)	.33 (2/6)	.17 (1/6)
Weak [.50 < w_M ≤ .75]	.29 (10/35)	.50 (3/6)	.33 (2/6)

(D) Strategy classification and its evidence in easy questions

	FM	FH	KI
Not Classified (N.C.) rate	0 (0/51)		
Classified rate	.02 (1/51)	0	.98 (50/51)
Very strong [.99 ≤ w_M]	0	0	.94 (47/50)
Strong [.95 < w_M ≤ .99]	0	0	.04 (2/50)
Positive [.75 < w_M ≤ .95]	1 (1/1)	0	0
Weak [.50 < w_M ≤ .75]	0	0	.02 (1/50)

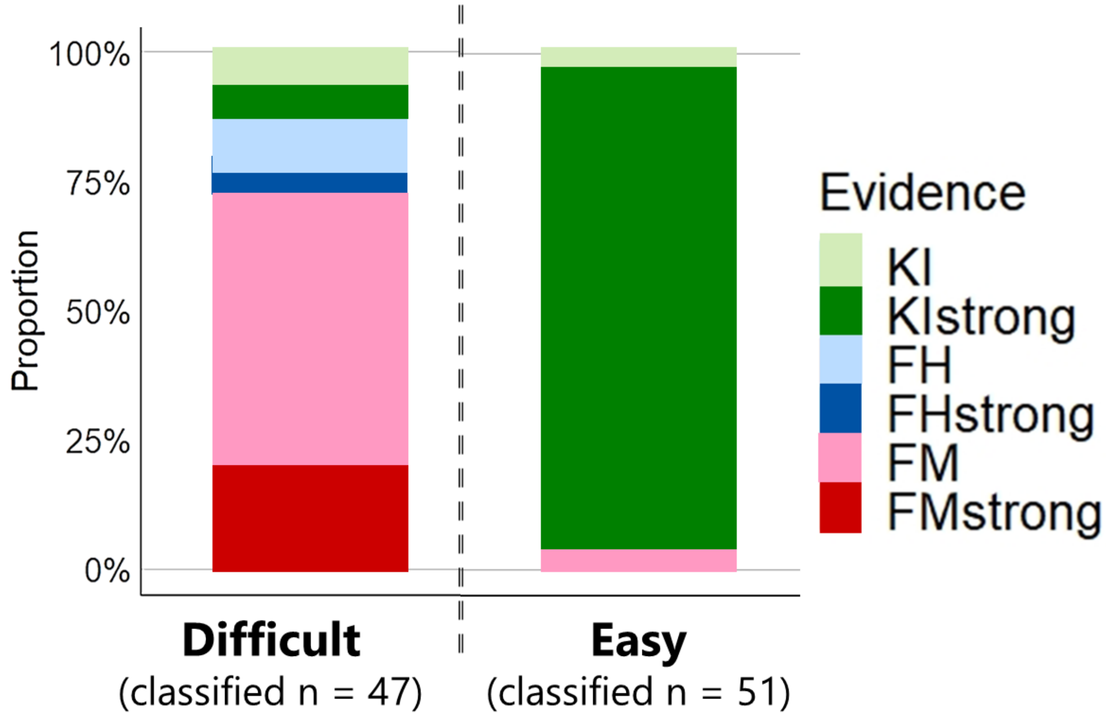


Fig. 13 Proportions of the strength of evidence for model selection in Study 3 (visualizing Table 3 (C) and (D)).

Note: FM: familiarity-matching model. FH: familiarity heuristic model. KI: knowledge-based inference model. The word "strong": "very strong" evidence (i.e., $.99 \leq w_M$) and "strong" evidence (i.e., $.95 < w_M \leq .99$).

5.3 Summary of Study 3

Study 3 replicated the main results obtained in Study 1. People's inferential tendencies which were observed in Study 1 could also be observed in this Study 3, even when the different objects were used. Through Studies 1, 2, and 3, it is confident that the key findings were highly robust and replicable.

Chapter 6 Study 4: Examinations of the adaptive use of inferential strategies in terms of accuracy and applicability (Data analyses and computer simulation)

How often can a certain inferential strategy be used (i.e., applicable) in tasks? If the strategy has ecological rationality but cannot be used (e.g., due to people's cognitive constraints such as decision threshold, as described below), it will not be effective for solving tasks and will not be regarded as a useful strategy. Especially in this thesis, if I intend to clarify the adaptive use of familiarity-matching, then I need to investigate familiarity-matching's applicability as well as its accuracy in a relationships-comparison task.

Generally speaking, people make inferences under many cognitive constraints (e.g., Binz, Gershman, Schulz, & Endres, 2020; Gigerenzer, 2008; Hoffrage & Reimer, 2004b; Lieder & Griffiths, 2020; Simon, 1990). One of the critical constraints that may affect the use of familiarity-based heuristics is people's *decision threshold*, or sensitivity of familiarity. For example, if the familiarity of an object X and that of another object Y are very similar to each other and people cannot discriminate between them (i.e., below their decision threshold, or not enough sensitivity), then they cannot apply familiarity-based heuristics and therefore will have to make random guessing. So far in this thesis, a "context" aspect of Simon's scissors has mainly focused on, such as a task structure and an environmental structure. In Study 4, on the other hand, a "cognitive" aspect of Simon's scissors should be more focused on, especially not only the way of solving tasks (e.g., choosing an alternative based on a similarity in familiarity) but also cognitive constraints such as decision threshold. For further understandings of the adaptive use of heuristics in a relationships-comparison task, it should be considered how often people have chances to use a certain heuristic (e.g., "$|FamQ - FamA| \neq |FamQ - FamB|$" for familiarity-matching [FM]; "$FamQ \neq medianFamQs$, and $FamA \neq FamB$" for familiarity heuristic [FH]; "know the correct country," or "do not know the country, and the number of recognized attributes for alt.A ≠ that for alt.B" for knowledge-based inference [KI]), that is, *applicability*.

In Study 4, I focused on people's decision threshold of familiarity in solving relationships-comparison tasks, and evaluated not only each strategy's *accuracy* (i.e., *correct rate*) but also each strategy's *applicability*. The definitions of these terms in this thesis are as follows:

Decision threshold (Sensitivity): The extent to which people can discriminate between two objects in terms of familiarity. Decision threshold has an inverse relationship with sensitivity. If one has a *low* decision threshold, then s/he can have many chances to discriminate two objects in terms of familiarity. In such a case, s/he has a *high* sensitivity, and vice versa. In short, "having a high (low) sensitivity" means "having a low (high) decision threshold." So, in order to simplify the descriptions about people's sensitivity, hereafter I will integrate these two concepts and will use one term, *decision threshold*, as in Study 1 and Study 3 (see also FM and FH model's assumption in Chapter 2).

Correct rate (*accuracy*): The extent to which people can make correct inferences by using the strategy, when the strategy is applicable. Here, the strategy's correct rate can be computed by "the number of correct inferences divided by the number of applicable cases."

Applicability: The extent to which a certain inferential strategy can be used in tasks that people work on. If there are many (little) chances where an inferential strategy can be used, then the strategy has a high (low) applicability. Here, a strategy's applicability can be computed by "the number of applicable cases divided by the number of all questions."

Study 4 consists of two parts: Analyses of behavioral data and computer simulations. The general purpose of Study 4 was to examine the adaptive nature of inferential strategies in terms not only of the correct rate but also of the applicability, in consideration of people's cognitive constraints (i.e., decision threshold). The main procedure in Study 4 was as follows.

Analyses of behavioral data (Study 4a): The analyses of the behavioral experiment data were conducted to examine which strategy could be adaptive within the three inferential models (i.e., familiarity-matching [FM], familiarity heuristic [FH], and knowledge-based inference [KI]), in terms of its applicability and correct rate in a relationships-comparison task. Using behavioral data provided from Study 3, I compared the applicability and correct rate between these three inferential models. Through the analyses, I could discuss why a certain heuristic, especially FM, would be an adaptive strategy.

Computer simulations (Study 4b): People's decision threshold was manipulated to investigate how their decision threshold would affect heuristics' applicability and correct rate by computer simulations based on *ACT-R* architecture (e.g., Anderson et al., 2004; Fechner, Pachur, & Schooler, 2019; Marewski & Schooler, 2011; Schooler & Hertwig, 2005). As a general procedure, I generated computer agents' familiarity for each object based on ACT-R. And then, heuristics' performances (i.e., computing the applicability and correct rate) for a relationships-comparison task were simulated with the assumption that these agents used FM or FH.

6.1 Study 4a: Confirming the adaptive nature of familiarity-matching based on behavioral data

So far, it was clarified that people often used familiarity-matching (FM) in a relationships-comparison task (Study 1) and its robustness (Study 3). It was also clarified that familiarity-matching would be an accurate strategy because it could reflect an environmental structure (i.e., ecological rationality; Study 2). However, if people have little or no chance to use an inferential strategy for current tasks, the strategy cannot be regarded as a really adaptive strategy even if the strategy has ecological rationality in a particular context. Specifically, by the definition of FM algorithm (see section 2.1), FM cannot be applied if "$|FamQ - FamA| = |FamQ - FamB|$" in a relationships-comparison task. If such cases are frequently observed, FM would be no longer regarded as "adaptive" even if FM has ecological rationality in a relationships-comparison task.

Then, the purpose of Study 4a was to compare which strategy (FM, FH, or KI) could be more adaptive in terms of each strategy's correct rate and applicability. Especially, using the behavioral data, the analyses were conducted to confirm whether FM (i.e., a frequently used and ecologically rational strategy in a relationships-comparison task) was more accurate and applicable than FH and KI.

6.1.1 Method

For analyzing the applicability in a task set and correct rates by using the strategies, the behavioral data provided from Study 3 were used. The reason why Study 3's data were used was that the materials in Study 3 were generated based on the materials in Study 2, in which we investigated ecological rationality of heuristics, and would reflect the real-world environmental structure well (i.e., not picked up by an experimenter's own criteria). So, it will be more appropriate to use Study 3's materials for examining the adaptive aspects of strategies than to use Study 1's materials.

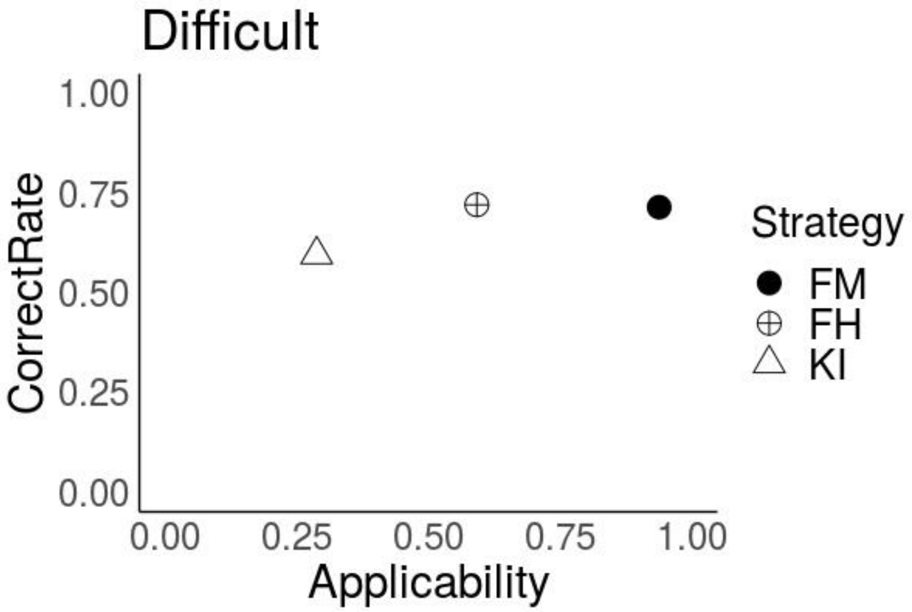
Then, for each difficulty level, the applicability and correct rate of three strategies were calculated for each participant. The mean of applicability and that of correct rate among participants were regarded as the strategy's applicability or correct rate, respectively. In calculating them, the decision thresholds that were estimated from the behavioral experiment were used for each participant (see section 3.2.1).

6.1.2 Results & Discussion

The applicability and correct rate were compared between FM, FH, and KI for each difficulty level. As to the correct rate (Fig. 14, y-axis), little differences were observed between strategies (Difficult: FM .72, FH .72, KI .60, all $ps > .06$; Easy: FM .98, FH .99, KI .98, all $ps > .10$; all one-way ANOVA, Holm adjusted in pairwise comparison). As to the applicability (Fig. 14, x-axis), on the other hand and interestingly, FM had the highest applicability out of three strategies both in difficult and easy questions (Difficult: FM .94, FH .59, KI .29, all $ps < .01$; Easy: FM .90, FH .78, KI .73, FM vs FH and FM vs KI $p < .01$, FH vs KI $p = .13$; all one-way ANOVA, Holm adjusted in pairwise comparison). It is indicated that, although FM's accuracy is not different so much from other inferential strategies, FM is more applicable than other strategies in a relationships-comparison task. These results were consistent with the results of Study 3 in terms of participants' decision threshold (see Table 3 (A)). Although only difficult questions were focused on since no participants were classified into FH in easy questions in Study 3, FM users had a higher decision threshold than FH users (FM 28.3, FH 16.7; not conducted a statistical test because of highly biased and small number of samples). That is, FM will generally be required a higher decision threshold (i.e., lower sensitivity) than FH, therefore FM can be regarded as a higher applicable heuristic in a relationships-comparison task (I will discuss the possible reasons for FM's high applicability in section 8.1).

In sum, although the correct rates are not so different between three strategies, FM (a most often used and ecologically rational strategy) is an adaptive strategy in a relationships-comparison task because it can have higher applicability in a task set.

Note that, the results of Study 4a may partially be inconsistent with those of behavioral experiments (Study 1 and Study 3) in terms of the "adaptive" use of inferential strategies. In Study 4a, the correct rates did not so differ between three strategies but the applicability was highest in FM both in difficult and easy questions. So, people may NOT need to shift the strategies depending on the difficulty levels, and using FM constantly may be better. I will discuss this issue in section 8.2.



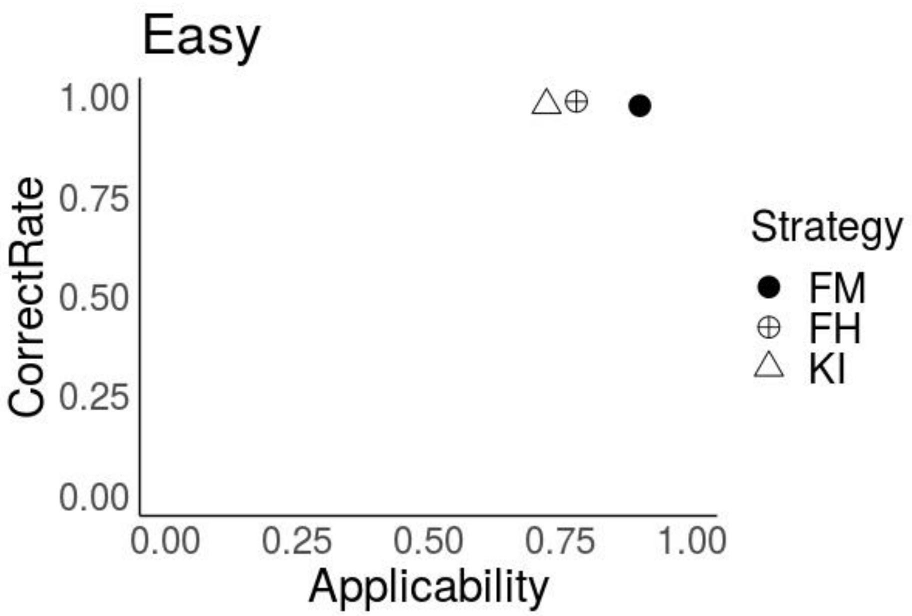


Fig. 14 The applicability and correct rate of familiarity-matching (FM), familiarity heuristic (FH), and knowledge-based inference (KI). These data were provided from Study 3. The x-axis and the y-axis denote the applicability in a task set and the correct rate in applicable cases, respectively. (Upper) Difficult questions. (Lower) Easy questions.

6.2 Study 4b: Examining effects of decision threshold through computer simulations

The applicability of familiarity-based heuristics will highly depend on people's subjective constraints, that is, the *decision threshold* for discriminating between similarities in familiarity (which has an inverse correlation with *sensitivity*, i.e., people with higher decision threshold are regarded as people with lower sensitivity, and vice versa). A decision threshold will differ from person to person. How applicable are heuristics under various decision thresholds? Especially, can familiarity-matching (FM) be more applicable than familiarity heuristic (FH) in a relationships-comparison task, for both people with high threshold and those with low threshold? To clarify these issues, heuristics' performances should be examined under conditions on high or low decision threshold. However, it is difficult to empirically manipulate people's decision threshold. In addition, it is also difficult to enormously increase the number of questions, the number of people who work on them, and the number of times people make inferences in a behavioral experiment. Then, in Study 4b, computer simulations were conducted to evaluate heuristics' performances by manipulating computer agents' decision threshold, assuming that many agents iterate to make many inferences.

The purpose of Study 4b was to theoretically examine the adaptive nature of FM (i.e., to examine whether FM's applicability, rather than its accuracy, would contribute to its adaptive nature). In simulating performances of heuristics, ACT-R architecture was adapted, as in many previous studies (e.g., Dimov & Marewski, 2017; Fechner et al., 2019; Fechner, Schooler, & Pachur, 2018; Honda 2020; Marewski & Mehlhorn, 2011; Marewski & Schooler, 2011; Schooler & Hertwig, 2005)

6.2.1 ACT-R model for investigating the adaptive nature of heuristics

Adaptive Control of Thought-Rational cognitive architecture, or *ACT-R* (e.g., Anderson et al., 2004) is a unified theory for explaining and applying various phenomena in cognitive science, not only for heuristic studies (e.g., associative recognition in Schneider & Anderson, 2012; multitasking in Salvucci & Taatgen, 2008). As to heuristic studies, Schooler and Hertwig (2005) modeled recognition and fluency heuristics using ACT-R, manipulating the probability and speed of retrieval of memories based on *activation* of memories (see also Fechner et al., 2019; Marewski et al., 2011). Generally speaking, ACT-R model can well capture dynamics of human memory, assuming that memories are stored as small units of information (*record* or *chunk*) (e.g., Fechner et al., 2019). ACT-R also assumes that the retrieval of memories depends on the extent of activation of the record, and that the baseline of activation depends on when (e.g., few days ago or long time ago) and how many times people have encountered the information.

Based on models in previous studies, especially in Schooler and Hertwig (2005), familiarity with objects can be assumed as activation of the record. It is because familiarity tends to decrease as longer time has passed since a person saw or heard the objects (i.e., an aspect of "when" people encountered the information), and also tends to be strongly and positively correlated with the number of encounters with the objects in the person's daily life (i.e., an aspect of "how many times" people encountered the information). In computer simulations, computer agents' familiarity (simulated familiarity) for each object was generated, regarding activation of the record as familiarity. Familiarity with objects can be generated based on the hit number in the real-world media, because people's familiarity will be highly correlated with the frequency of encounter (see Study 2). Then performances of familiarity-based heuristics in hypothetical relationships-comparison tasks are simulated, assuming that computer agents constantly use FM or FH in all possible questions (see Study 2).

The outline of the computer simulations in Study 4b is as follows. First, using 125 objects in Study 3 (i.e., 100 cities and 25 countries), I confirmed whether the procedure of simulations would be appropriate by comparing the simulated familiarity with the actual familiarity ratings provided by participants of Study 3. Next, using another 125 objects (not used in Study 3), I generated familiarity with each object in the same way. Then, heuristics' performances (i.e., correct rate and applicability) for a relationships-comparison task were simulated with manipulations of agents' parameter of decision threshold. By conducting these simulations, it was expected to obtain theoretical understandings of whether FM's applicability would really contribute to its adaptive nature (which is the finding provided from the analyses of behavioral data; see Study 4a).

6.2.2 Method

Based on the procedure in previous studies (e.g., Fechner et al., 2019; Marewski et al., 2011; Schooler & Hertwig, 2005), computer simulations were conducted in order to examine the adaptive use of heuristics in the following four steps.

Step 1: Simulating the probability of encounter, $P(i)$*, with object* i: First, the probability that a computer agent will encounter a particular object was simulated. Generally, the more frequently objects appear in the real world, the higher the probability of encounter with them will be. So, to simulate each agent's probability of encounter with objects (i.e., city names and country names) in the agent's life, the real-world data were collected according to the way of Schooler and Hertwig (2005) (see also Study 2). As an index for a frequency of city names' and country names' appearing in the real world, the hit numbers in an online-newspaper database, *Maisaku* (provided by a famous Japanese newspaper, Mainichi-shimbun)[xiv] were used. I set the time window to 5,000 days, and counted the frequency of appearance (i.e., hit number) in the database for the 5,000 days (from 06/10/2020 to 10/02/2006). Because these hit numbers were extremely skewed, I transformed the original hit numbers into log scale for each city or for each county. Then, using these log-transformed hit numbers, I defined the probability of encounter, $P(i)$, for encountering object i as:

$$P(i) = \frac{HitNumber_i}{\max(HitNumber)} * C \quad (4)$$

where $HitNumber_i$ is the hit number of object i in the database, and $\max(HitNumber)$ is the largest hit numbers in the database (475,927 in Step 3 [odd-ranked data]; 32,725 in Step 4 [even-ranked data]; see Supplementary material 5), and C is a scaling parameter for adjusting the max probability. I set $C = .90$ for country names and $C = .80$ for city names. That is, it was assumed that the most frequently appeared county name and city name were encountered with the probability of .90 and of .80 in each day, respectively. Note that, the max probability of encounter with city names was set nearly .80 in previous studies (e.g., "Berlin" with the probability .73 in Schooler and Hertwig (2005)), but it was considered that country names were more likely to be seen and heard than city names in people's daily lives, thus I set the max probability of encounter with country names as .90 for in this thesis (see also "Data from the measurement of familiarity task" in Supplementary material 2).

Step 2: Calculating familiarity with objects based on the probability of encounter. Next, I calculated familiarity with objects for each agent. Since familiarity with objects is strongly correlated with how often people saw or heard the objects in their daily lives, I simulated when and how often an agent encountered object i in a certain time window (5,000 days in this case), based on the probability of encounter $P(i)$ provided in Step 1. Specifically, it was assumed that an agent's encountered object i at the day t, thus encountered it $5{,}000 * P(i)$ times in average within 5,000 days. The simulated data for encountering object i were regarded as the agent's historical record (i.e., storage of the agent's memories). This record could be regarded as the agent's subjective memory experiences. Using the record simulated by the above procedure, I calculated the familiarity of object i based on the following ACT-R model for each agent (e.g., Schooler & Hertwig, 2005):

$$Activation_i = Baseline_i + \sum_{j=1} Spreading_{ji} \quad (5)$$

$$Baseline_i = \ln\left(\sum_{j=1}^{n} t_j^d\right) \tag{6}$$

where $Activation_i$ and $Baseline_i$ denote the activation of memories and the baseline of the activation, respectively, for object i. $Spreading_{ji}$ denotes the spreading activation (e.g., contextual effects such as agents' mood), which is ignored in the current simulations following Schooler and Hertwig (2005). The baseline of activation is represented by (i) how many times, n, an agent encountered object i in the time window, (ii) the jth encounter occurred the day t_j in the time window, and (iii) the amount of decay (i.e., forgetting), d, in the agent's memories. I set d to the typical value, − 0.5 (e.g., Schooler & Hertwig, 2005). Taken together, familiarity with object i is regarded as $Activation_i = Baseline_i = \ln\left(\sum_{j=1}^{n} t_j^{-0.5}\right)$ in the current model. The examples about familiarity with object i are as follows. Imagine situations where two agents, X and Y, encountered object i total 4 times in the past: If an agent X encountered it 100, 400 and 600 days ago, then X's familiarity with i is "$\ln(100^{-0.5} + 400^{-0.5} + 600^{-0.5}) = -1.66$." On the other hand, if an agent Y encountered it 10, 50, 100 and 200 days ago, then Y's familiarity with i is "$\ln(10^{-0.5} + 100^{-0.5} + 200^{-0.5}) = -0.46$." As shown in these example, the more often and the more recently an agent encountered object i, the higher an agent's familiarity with it would become in ACT-R model. I transformed the familiarity into z-score for each agent, and then regarded the mean of z-scored familiarity in 51 agents as simulated familiarity with object i.

Step 3: Conducting the check of consistency between behavioral and simulated data: To confirm whether the procedure of simulating familiarity could be appropriate, I compared the simulated familiarity with actual familiarity ratings provided from participants in Study 3. Using 125 objects in Study 3 (i.e., 25 out of 50 countries in the highest population size with odd rank [ranked 1st, 3rd, ⋯, 49th], and the top 4 cities in the population size from each country [25 countries * 4 cities = 100 cities]), I generated familiarity with objects according to Steps 1 and 2. In this check, I set the number of agents to 51, which was the same number as the number of participants in Study 3. Both for simulated data and for behavioral data, I calculated the mean familiarity with each object. The hit numbers of odd-ranked 125 objects can be found in Supplementary material 5 (upper table). The correlation between the simulated familiarity and familiarity in the behavioral experiment (Study 3) is shown in Fig. 15 in the next "Result & Discussion" section.

Step 4: Evaluating heuristics' performances: After confirming the appropriateness of this procedure, I next used another 125 objects (not used in Study 3) and generated familiarity with objects in the same way. Specifically, as materials for the current simulations, I used 25 out of 50 countries in the highest population size with even rank (ranked 2nd, 4th, ⋯, 50th), and the top 4 cities in the population size from each country (25 countries * 4 cities = 100 cities). The hit numbers of even-ranked 125 objects can be found in Supplementary material 5 (lower table). Then, the computer simulations were conducted to evaluate the applicability and correct rate of heuristics (i.e., FM and FH), with the following parameters. I simulated 10,000 agents' inferences based on FM and FH algorithms (specifically, the number of agents was set to 100, and each agent solved all possible alternative pairs [i.e., correct alternative 100 patterns * false alternative 24 patterns = 2,400 questions] for 100 times). In computing the applicability and correct rate, agents' decision thresholds (simulated threshold) were set to eight patterns: 0.00 (i.e., the highest sensitivity), 0.01, 0.05, 0.10, 0.20, 0.50, 0.90, and 1.00 (i.e., the lowest sensitivity). It was assumed that, if the condition for the decision threshold is not satisfied (i.e., "$||\mathrm{FamQ} - \mathrm{FamA}| - |\mathrm{FamQ} - \mathrm{FamB}|| < \text{decision threshold}$" for FM; and "$\min\{|FamQ - medianFamQs|, |FamA - FamB|\}$" < decision threshold" for FH) then the agent made random guessing and a correct inference with the probability of .50.

6.2.3 Results & Discussion

This section will show the results of the computer simulations in terms of the manipulation check (Step 3) and heuristics' performances (Step 4). Then, FM's possible advantages for a relationships-comparison task are discussed.

The result of the check of consistency between behavioral and simulated data (Step 3): Fig. 15 shows a correlation between familiarity simulated in Study 4b and familiarity provided from a behavioral experiment in Study 3. Before simulating agents' inferential performances, I checked whether the simulated familiarity well reflected actual human data provided in the behavioral experiment. The correlation coefficient ρ between them was .83 (S = 56,767, $p < .001$, Spearman's rank correlation; Fig. 15). This high correlation indicates that the procedure of simulation could well reflect the actual people's familiarity and would be able to be regarded as an appropriate procedure.

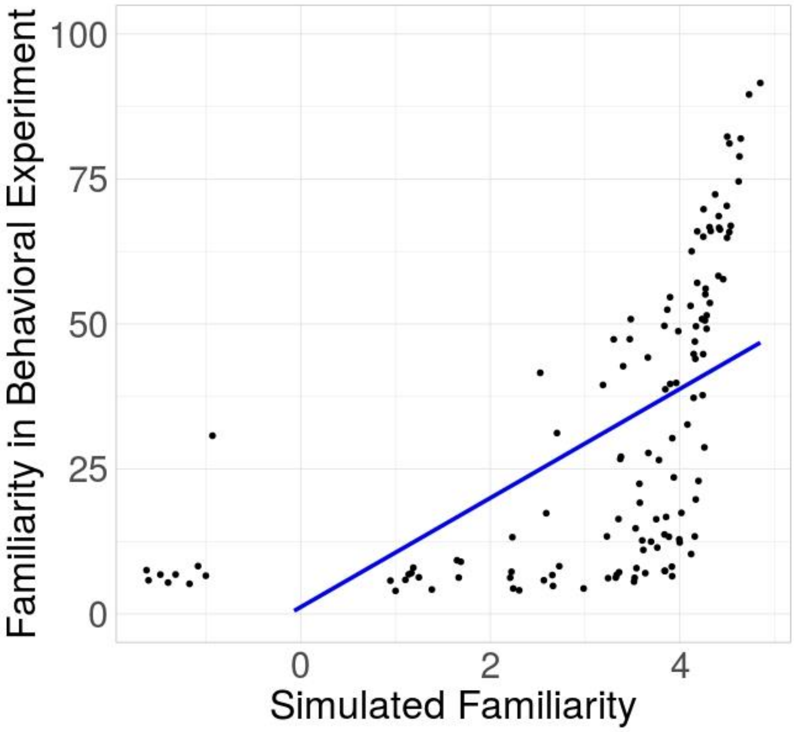


Fig. 15 Results of the check of consistency between behavioral and simulated data (Step 3). Correlation between the simulated familiarity (x-axis) and familiarity ratings provided from behavioral experiment (y-axis). In the current simulations, the decay parameter, d, was set to –0.5 (typical value; e.g., Schooler & Hertwig, 2005). Each point denotes each object. Blue line denotes a linear regression line (Spearman's rank correlation $\rho = .83$, $p < .001$).

The result of heuristics' performances (Step 4): Next, the applicability and correct rate of FM or FH were evaluated with manipulating computer agents' decision threshold. Specifically, based on the findings of Study 4a, it was predicted that FM's higher applicability would contribute to its adaptive nature in a relationships-comparison task, and I confirmed this prediction by conducting computer simulations. As in Study 4a, FM's correct rate were not so much different from FH's correct rate (Fig. 16 (A)). On the other hand, FM had constantly higher applicability than FH, regardless of agents' decision threshold (Fig. 16 (B)).

These results indicate that, regardless of the extent of people's decision threshold for discriminating between similarities in familiarity, FM is easier for people to apply than FH and therefore FM has higher adaptivity in a relationships-comparison task. The results may also imply that people are likely to use a more applicable strategy (i.e., FM) under their cognitive constraints. Based on the results of Study 4b's computer simulations, I could reveal the adaptive nature of FM theoretically.

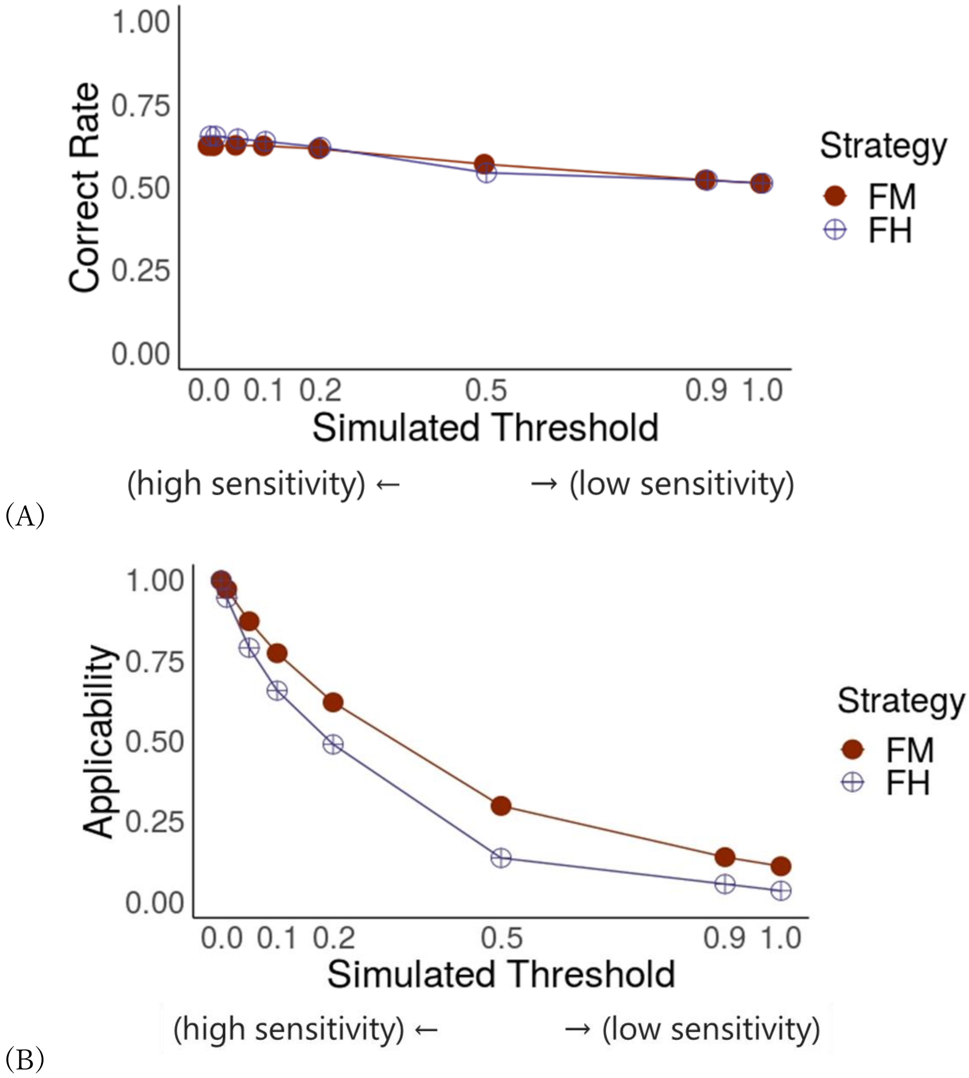


Fig. 16 Results of heuristics' performances (Step 4 in Study 4b). The x-axis denotes simulated agents' decision threshold (inversely correlated with *sensitivity* to discriminate between similarities in familiarity). In the parameter settings, agents' decision thresholds were set to 0.00 (i.e., the highest sensitivity), 0.01, 0.05, 0.10, 0.20, 0.50, 0.90, and 1.00 (i.e., the lowest sensitivity), shown by each data point. (A) The correct rate of heuristics (i.e., familiarity-matching [FM] and familiarity heuristic [FH]). (B) The applicability of them.

6.3 Summary of Study 4

Familiarity-matching (FM) is often used and can have ecological rationality in a relationships-comparison task, but if FM cannot be applied to many cases, it will not be regarded as a useful strategy. The use of heuristics will highly depend not only on a context aspect (e.g., task structure) but also on a cognitive aspect (e.g., cognitive constraints), especially people's decision threshold, or *sensitivity*, for discriminating between objects in terms of familiarity. Then, Study 4 examined inferential strategies' *applicability* and *correct rate* in a relationships-comparison task. The results of Study 4a (analyses of behavioral data) showed that, although the correct rates did not so differ between strategies, FM was more applicable than familiarity heuristic (FH) and knowledge-based inference (KI). Furthermore, the results of Study 4b (computer simulations in which the parameter of agents' decision threshold was manipulated) indicated that FM was an applicable strategy regardless of people's decision threshold. Taken together, in addition to the use of FM (Study 1 and Study 3) and the accuracy of FM (Study 2), FM is a highly applicable strategy for a relationships-comparison task in difficult questions.

Chapter 7 Study 5: familiarity-matching in a daily context (Behavioral experiment)

From Study 1 to Study 4, I examined the use of familiarity-matching (FM) only using tasks in which people were asked to make inferences based on their general knowledge retrieved from memory about geographical features. As in Gigerenzer and his colleagues' works, such general knowledge can be seen through media in the real world, and I focused on a specific environmental structure according to the framework of adaptive toolbox (e.g., Gigerenzer et al., 1999). So, it still remains unclear whether people (not only university students) will tend to use a strategy like FM even in preferential, not inferential, choices such as in daily situations. If a task structure does not change (i.e., two alternatives and one object in a question are presented, and people are required to compare two dyad relationships), then do people use FM even in preferential tasks? In Study 5, as an additional study, I used a task that had the same task structure as that used in Studies 1~4 but did not ask people's general knowledge. Especially, I focused on a possible daily situation involving consumer choice behaviors (e.g., del Campo, Pauser, Steiner, & Vetschera, 2016; Hauser, 2014; Hilbig, 2014; Thoma & Williams, 2013), and aimed to investigate whether people used FM strategy in such a context. Note that, therefore, investigating ecological rationality of heuristics was not the purpose of Study 5.

I expect that, in consumer choice behaviors, people may use FM to consider the relationships between the familiarity of each alternative and that of a third, non-alternative object. If the familiarity of an alternative is more similar to familiarity of a third object in a consumer choice, then people may sometimes choose the less familiar alternative. For example, consider a situation in which a Japanese person wants to purchase soy sauce. Since the person uses a type of soy sauce daily, it will be highly familiar to her/him. S/he finds two types of soy sauces in a store: One is made by a very familiar company and the other by an unfamiliar one. In this situation, FM predicts that the person will choose a soy sauce made by the familiar company because the soy sauce is similarly familiar to her/him. On the other hand, consider another situation; a Japanese person intends to purchase nuoc mam (a kind of fish sauce which is sometimes used in Vietnamese cuisine). Since the person seldom uses nuoc mam, it is very unfamiliar to her/him. S/he finds two types of nuoc mam at the store: One is made by a highly familiar company and the other by an unfamiliar one. In this situation, FM predicts that s/he may choose the nuoc mam by the unfamiliar company because nuoc mam, which is an item s/he intends to purchase, is unfamiliar to her/his. Like the above examples, people may consider a similarity in familiarity among an item and companies in their daily lives.

In such a relationships-comparison situation, FM can predict cases not only of a buyer choosing the more familiar alternative but also of a buyer choosing the less familiar one. In fact, based on the survey of sales rankings, I found that an item's rank could vary, depending on the company that produced it. For example, a yogurt made by Morinaga Milk (a very famous company in Japan) is ranked among best-selling yogurts; while a yogurt made by Marusan (a less famous company that sells many yogurts) is out of the sales ranking. However, a soy milk product (generally less familiar than yogurt) made by Marusan is among the most popular in Japan; while a soy milk product made by Morinaga Milk is not in the ranking. These ranking data imply that people may consider the relationships between the companies (two alternatives) and the item (a third, non-alternative object) in their consumer behaviors[xv].

The main procedure of Study 5 was as follows. In order to obtain the first evidence for the strategy that people use in a daily preferential context, I conducted a behavioral experiment using materials for consumer choices (i.e., item names and company names). Based on the results, I calculated the accordance rate between the prediction from FM (or from familiarity heuristic [FH]) and participants' choice patterns, and then compared the accordance rate of FM with that of FH, for each participant. Moreover, I also examined to what extent the accordance rate of FM's prediction was correlated with people's risk attitudes or with people's concerning levels about the production area.

7.1 Participants

One-hundred and twenty people (M_{age} = 44.3, SD_{age} = 8.21, and 61 participants were female) have participated in Study 5. To examine ordinary people's usage of FM, I recruited participants with a wide range in age (each 30 participants were recruited from each 30-39, 40-49, 50-59, and 60-69 years old).

7.2 Tasks, materials & procedure

I conducted the binary choice task (consumer choice version) and the measurement of familiarity task via an online questionnaire form. Participants answered consumer choice questions such as "if you want to buy item Q, which do you want to buy, made by company A, or made by company B?" The list of questions can be found in Supplementary material 6. I used seven categories of item (e.g., alcohol, seasoning, tea, etc.). Each category had both familiar and unfamiliar items and companies. I made two types of situations in one category: One situation was that participants intended to buy a familiar item and were asked to choose the item made by a familiar company or an unfamiliar company. Another situation was that they intended to buy an unfamiliar item and were asked to choose the item made by a familiar company or an unfamiliar company. Therefore, participants answered 14 questions (i.e., 7 categories * 2 situations). Furthermore, I inserted eleven lottery-choice tasks, such as "which alternative do you choose, 'sure of 5,000 yen,' or 'x % of 10,000 yen but 100 – x % of 0 yen?'" (x were 11 values: 1, 10, 20, ⋯, 90, 99) among the consumer choice questions, in order to measure participants' levels of risk attitude. The order of questions and alternatives was randomized.

After the binary choice task, participants answered the measurement of familiarity task. This task was conducted with the same procedure as in Study 1.

At the end of the experiment, participants were asked to rate how much they cared, in their daily lives, about the place or company where an item was made, using a visual analog scale (0 [do not care at all] – 100 [care very much]).

7.3 Results & Discussion

First, for a manipulation check, I compared familiarity of "familiar items" with that of "unfamiliar items." I converted the familiarity of objects into z-scores for each participant and then calculated the mean of familiarity. As a result, the familiarity of "familiar items" was significantly higher than that of "unfamiliar items" (mean: 0.75 and -0.38, respectively; $V = 7.3 * 10^3$, $p < .001$, $r = 0.61$, Wilcoxon rank sum test). The familiarity of "familiar companies" was also higher than that of "unfamiliar companies" (mean: 0.63 and -1.00, respectively; $V = 7.2 * 10^3$, $p < .001$, $r = 0.61$, Wilcoxon rank sum test). These results indicated that my experimental materials were well identified as expected.

Then, I calculated the accordance rates of FM and FH for each participant, to examine the extent to which the observed participants' choices accorded with the prediction from FM or FH in the 14 questions (as to FM and FH algorithms, see section 2.1 and 2.2; however, I did not consider participants' decision threshold in Study 5). The mean accordance rate of

FM was .61, which was significantly higher than that of FH, at .40 (the solid vertical line in the upper and lower panels of Fig. 17, respectively; $V = 5.1 * 10^3$, $p < .001$, $r = 0.49$, Wilcoxon rank sum test).

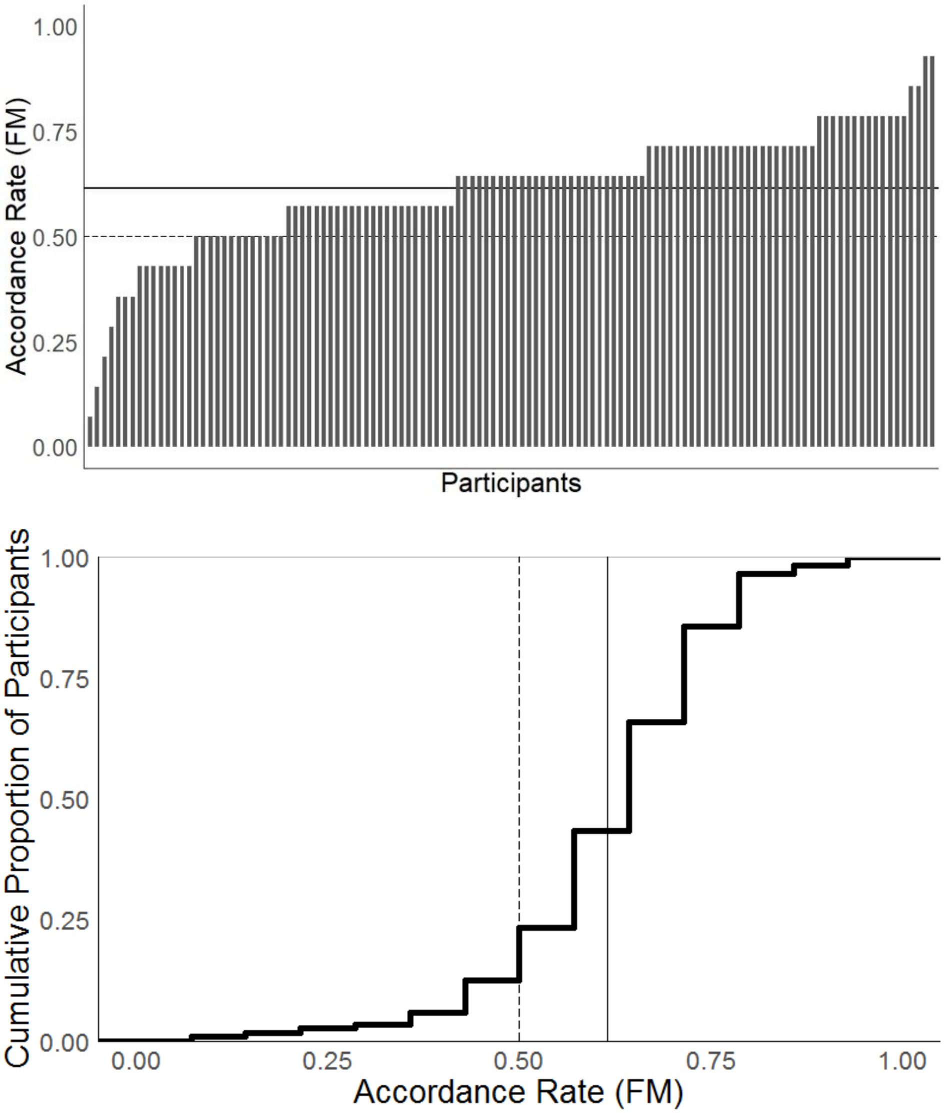
For further analyses, I examined the correlations between the accordance rates and participants' levels of risk attitude, as well as between the accordance rates and participants' levels of concerning for production areas. As described earlier, participants answered the lottery choice tasks in the behavioral experiment. So, I could obtain an index for the level of risk attitude for each participant, based on the point at which the participant shifted to choose the risky alternative. For example, if a person chose 'sure of 5,000 yen' when x was from 1 to 70 but chose 'x % of 10,000 yen but 100 – x % of 0 yen?' (i.e., risky alternative) when x was from 80 to 99, then I defined the parson's level of risk attitude as "3" because "x = 80" was the 3rd highest value among the 11 values of x (hereafter, I call this "Risk Seek Level"). In addition, at the end of the experiment, participants were rated how much they cared about the production areas in their daily lives. So, I could also obtain an index for the level of concerning about production areas or companies for each participant, based on the participant's rating score (i.e., 0 [do not care at all] – 100 [care very much]) (hereafter, I call this "Concerning Level"). However, neither of these indexes correlated with the accordance rate of FM or FH (Fig. 18: Spearman's rank correlations between "risk attitude and FM accordance rate," "concern level and FM accordance rate," "risk attitude and FH accordance rate," and "concern level and FH accordance rate" were "$r = -.02$, $p = .84$," "$r = .01$, $p = .75$," "$r = -.01$, $p = .91$," and "$r = -.03$, $p = .89$," respectively). These results indicate that participants who were more likely to be risk-seeking or to be concerned about the production area do not show a particular tendency to use either the FM or the FH strategy.

I believe that people can sometimes apply a decision strategy like familiarity-matching outside the laboratory, and therefore, FM can show adaptive aspects for relationships-comparison situations when people do not have sufficient knowledge. Furthermore, I also believe that the findings in Study 5 will contribute to the consumer choice theory. Generally, people often choose the more familiar out of two alternatives, not only in inferential situations but also in preferential situations (e.g., a consumer choice; e.g., Hilbig, 2014; Michalkiewicz, Arden & Erdfelder, 2018). However, if people use a judgmental strategy of comparing between two dyadic relationships like FM (i.e., "item Q and company A" and "item Q and company B"), then they will choose the *less* familiar alternative when a third object (item Q) is unfamiliar. Thus, it may be possible to explain a "boundary" in the choice of a familiar/unfamiliar alternative (i.e., when and in what situations people are likely to buy the more/less familiar alternative). I believe that, by investigating consumer choices in a structure of a relationships-comparison task (i.e., not a situation of a simple comparison of alternatives), new findings not only about inferences but also about preferences can be obtained. Because the strategy of FM could significantly predict participants' choice patterns better than the strategy of FH, it is expected that people will really choose items based on a similarity in familiarity even in their daily preferential choices.

Finally, I note two unclear questions related to Study 5. First, were there some confounding factors that would affect the observed results? Of course, not only people's familiarity with objects but also other information or their further knowledge (e.g., a company's brand power; whether the item was domestic products or foreign products) might affect their choice behaviors. If these factors are considered, better explanations about the strategy that people used in a preferential context may be able to be provided.[xvi] Although other confounding factors might exist, I believe that I could obtain the first evidence for people's use of FM also in preferential choices, which is an important contribution of Study 5.

Second, what made people use FM in preferential choices? I have no choice but to speculate this issue. One possible interpretation is the explanation based both on an environmental structure and on people's experiences in the real world. Sometimes, people may experience cases like the following: If people face familiar objects (i.e., objects that will be frequently seen or heard in the real world), then generally they can deal with them or can obtain a successful result by choosing a familiar alternative, and vice versa. Consider the following two questions: "Which do you want to buy to take a safeguard measure for influenza virus, an alcohol-based sanitizer or a sodium hypochlorite-based sanitizer?", and "Which do you want to buy to take a safeguard measure for rotavirus, an alcohol-based sanitizer or a sodium hypochlorite-based sanitizer?" Generally, people are more familiar with influenza virus or alcohol than rotavirus or sodium hypochlorite, respectively. Usually, a familiar virus can be sanitized by the familiar sanitizer (i.e., for influenza virus, buy alcohol; in the former case). However, people may think that an unfamiliar virus is strong, and a familiar sanitizer may not be effective for it (i.e., for rotavirus, buy sodium hypochlorite; in the latter case). In fact, it is believed that rotavirus has a greater tolerance for alcohol than influenza virus, and sodium hypochlorite is more effective for rotavirus than alcohol[xvii]. Like such cases, people may learn that it will be likely to be successful if they match objects based on the similarity in familiarity, from their daily experiences. Such experiences may make people use FM in preferential choices.

(A)



1.00
0.75
0.50
0.25
0.00
Accordance Rate (FM)
Participants
Cumulative Proportion of Participants
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Accordance Rate (FM)

(B)

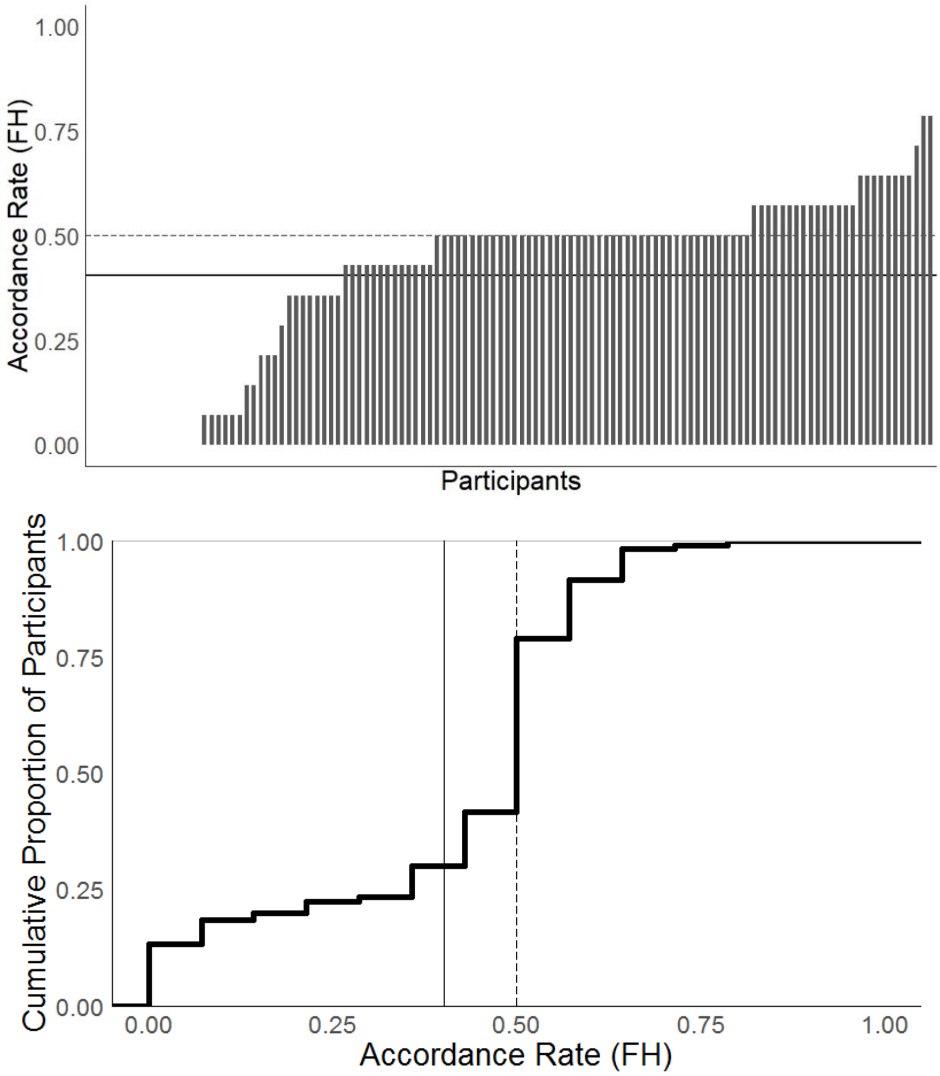
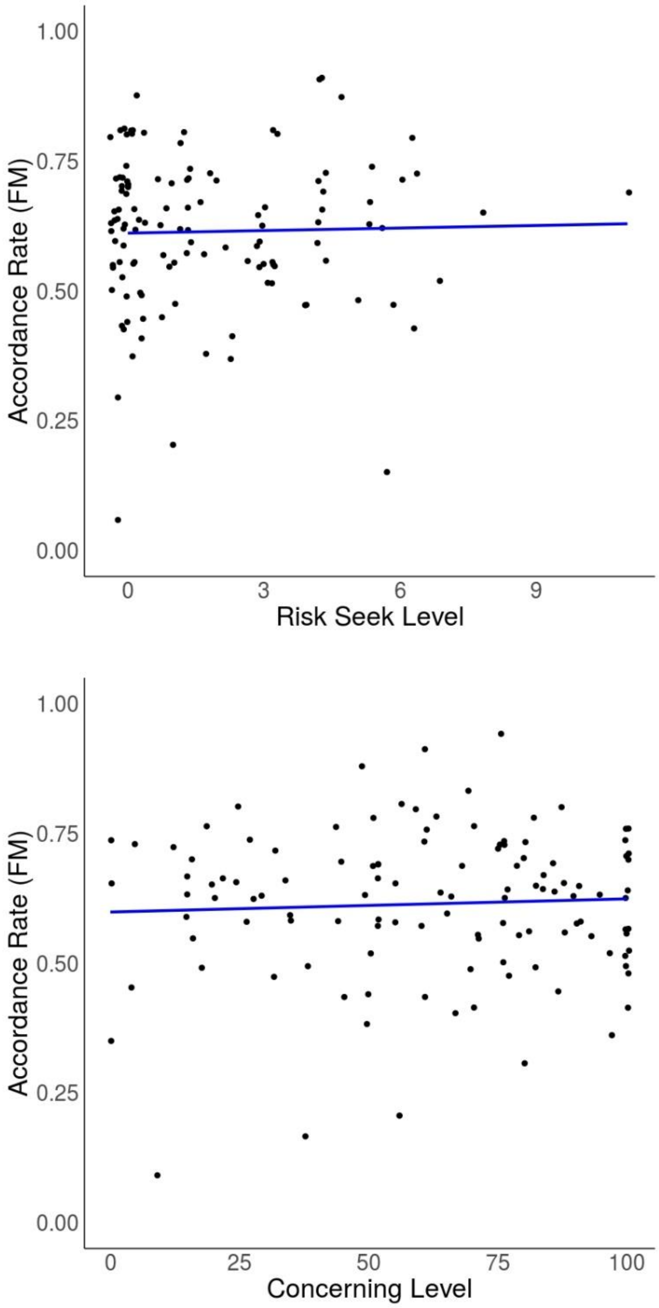


Fig. 17　Results in Study 5 as to (A) familiarity-matching and (B) familiarity heuristic. Both in (A) and (B), the upper panel shows individual data (N = 120) and the lower panel shows the empirical cumulative distribution functions for accordance rates (i.e., the extent to which participants' observed choices accorded with the predictions from FM or FH in the 14 questions). The dotted line denotes the chance level (.50), and the solid line denotes the mean accordance rates (FM: .61; FH: .40).

(A)



1.00
0.75
0.50
0.25
0.00
Accordance Rate (FM)
0
3
6
9
Risk Seek Level
1.00
0.75
0.50
0.25
0.00
Accordance Rate (FM)
0
25
50
75
100
Concerning Level

(B)

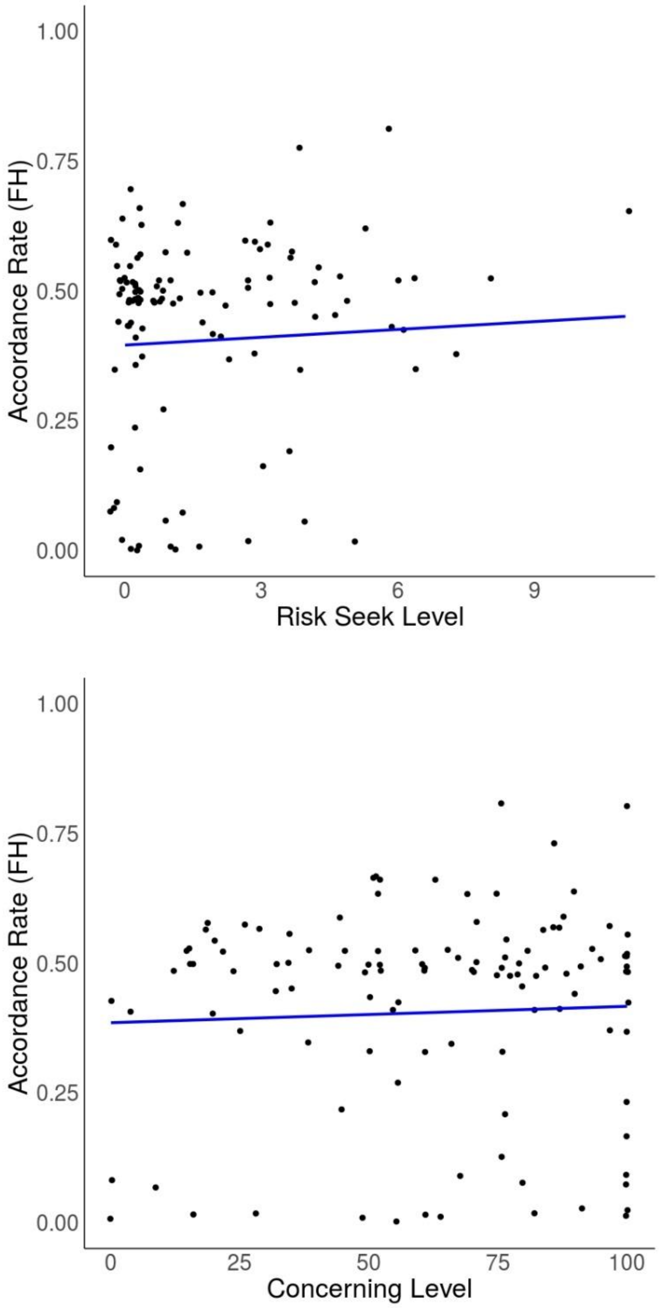


Fig. 18 Correlations between “Risk Seek Level” or “Concerning Level” and accordance rates from (A) familiarity-matching and (B) familiarity heuristic. Spearman’s rank correlation coefficients, from the top panel to the bottom panel, are “r = -.02, p = .84,” “r = .01, p = .75,” “r = -.01, p = .91,” and “r = -.03, p = .89,” respectively.

7.4 Summary of Study 5

To obtain the first evidence of the strategy that people use for a relationships-comparison task in a daily context. I focused on consumer choices. As a result, people's choices in the behavioral experiment highly accorded with the prediction from familiarity-matching (FM) model. The findings obtained in Study 5 suggest that people will use the judgmental strategy like FM in a structure of relationships-comparison task, not only in an inferential context but also in a daily context.

Chapter 8 General discussion

The Human system has the *adaptive toolbox*. Gigerenzer and his colleagues have argued that there are various strategies in this toolbox, and depending on an environment (i.e., the real world that can be seen or heard through media and documents), people intuitively select the most useful (i.e., adaptive) strategy among them in terms of accuracy and applicability. As Herbert A. Simon and many cognitive scientists have pointed out, human intelligence should be understood not only from “cognition” aspects but also from “context” aspects (i.e., a metaphor of Simon’s scissors). Generally, people use simple heuristics under various cognitive constraints such as limited their knowledge and decision threshold. However, when the structure of heuristics (i.e., “cognition”) matches well with that of environments (i.e., “context”), the heuristics can often work well (i.e., ecologically rational).

In this thesis, I investigated people’s adaptive use of heuristics according to Gigerenzer’s adaptive toolbox framework. Specifically, I focused on a *task structure* of a binary choice (i.e., the location in a problem statement where objects are presented, and the computation that a person is required to solve it) as a new context aspect. In previous studies on heuristics within the adaptive toolbox framework, only one task structure was used: Two alternatives were presented, and a simple comparison between these two alternatives was required (e.g., population estimation task). In such a task, people could exploit an environmental structure wherein people’s familiarity and a criterion (e.g., familiarity with a city and population size of the city) was highly correlated with each other. However, I predicted as follows: If a task structure of a new task differed from the structure of a previous task, then an environmental structure that people might exploit would differ, and adaptive heuristics for the new task would also differ. If people made adaptive inferences depending on an environment, do they use a new heuristic in the new task structure?

In order to investigate this issue according to the adaptive toolbox framework, I proposed a new task, *relationships-comparison task*, and then constructed three inferential models to describe human inferences for this task, *familiarity-matching* (FM; newly proposed in this thesis), *familiarity heuristic* (FH), and *knowledge-based inference* (KI). Through five studies, I examined the strategy that people used in a relationships-comparison task, and the accuracy and applicability of the strategy. Now, I summarized the focuses, main purposes, approaches, and main findings in these studies in Table 4.

Table 4 Summary of this thesis

	Focus	Main purpose	Approach	Main finding
Study 1	The strategy that people use in a relationships-comparison task	Examining what strategies can be used	Behavioral experiment	Familiarity-matching (FM) is often used in a relationships-comparison task
Study 2	Accuracy (in terms of ecological rationality)	Examining an environmental structure	Behavioral experiment	FM can reflect the real-world environmental structure well
			Analyses of the real-world data	
			Computer simulations	FM can lead correct inferences
Study 3	Replication of Study 1	Confirming the robustness of Study 1's findings	Behavioral experiment (using Study 2's materials)	The use of FM in a relationships-comparison task is robust
Study 4a	Applicability (under cognitive constraints)	Examining which strategy is more accurate and applicable	Analyses of the behavioral data	FM has the higher applicability than other inferential strategies
Study 4b		Examining Study 4a's purposes under various decision threshold (sensitivity)	Computer simulations	FM can have the higher applicability regardless of people's decision threshold
Study 5	The strategy that people use in preferential choices	Examining whether FM is generally used in a daily life	Behavioral experiment	FM can be used even in a daily context (consumer choices)

8.1 Why does familiarity-matching have higher applicability? Possible reasons based on the comparison with familiarity heuristic

Why can FM be more applicable than FH in a relationships-comparison task (see the results of Study 4)? One possible explanation is that using FM will NOT require lower decision threshold (i.e., higher sensitivity). In the assumption of FM model (see section 2.1), FM can be applied when "$||\mathrm{FamQ} - \mathrm{FamA}| - |\mathrm{FamQ} - \mathrm{FamB}||$ > decision threshold." That is, familiarity which is compared with a person's decision threshold is only "$||\mathrm{FamQ} - \mathrm{FamA}| - |\mathrm{FamQ} - \mathrm{FamB}||$" in FM, and therefore FM is more likely to be applicable even if the person has a relatively higher decision threshold. In the assumption of FH model, (see section 2.2), on the other hand, FH can be applied when "$|FamQ - medianFamQs| >$ decision threshold and $|FamA - FamB| >$ decision threshold." That is, there are two terms of familiarity to be compared with decision threshold: "$|FamQ - medianFamQs|$" and "$|FamA - FamB|$." In FH, a person's decision threshold should be lower than both of the two (i.e., required "min{$|FamQ - medianFamQs|$, $|FamA - FamB|$}" > decision threshold"). Because of this, a lower decision threshold will be required for using FH. To confirm these predictions, using the simulation data in Study 4b, I conducted additional analyses on distributions of distances of familiarity which is compared with the decision threshold (i.e., "$||\mathrm{FamQ} - \mathrm{FamA}| - |\mathrm{FamQ} - \mathrm{FamB}||$" in FM; and "min{$|FamQ - medianFamQs|$, $|FamA - FamB|$}" in FH). If the distance was larger, heuristics could be applied even when a person's decision threshold was higher. As shown in Fig. 19, the distance in FM was larger than that in FH (Median: FM 0.30, FH 0.24; $W = 3.3 * 10^6$, $p < .001$, $r =$ 0.12, Wilcoxon rank sum test), indicating that FM is more likely to be used than FH even if people have a higher decision threshold (i.e., a lower sensitivity was required to use FM than to use FH). Therefore, FM can have high applicability for a relationships-comparison task.

One may argue that the above comparison of applicability is unfair in terms of the number of familiarity which is compared with the decision threshold. In FM, a decision threshold is compared with only one term: "$||\mathrm{FamQ} - \mathrm{FamA}| - |\mathrm{FamQ} - \mathrm{FamB}||$" In FH, on the other hand, it is compared with two terms: "$|FamQ - medianFamQs|$" and "$|FamA - FamB|$." Based on this consideration, I also conducted further analyses. I constructed another FH model which assumes that people constantly choose the familiar alternative (called *naïve FH*, in which a decision threshold is compared with only one term) and compared its performances with FM and FH through Study 4b's simulations. As a result, I confirmed that FM would be more adaptive heuristic than naïve FH in terms of its correct rate. For the detailed explanations and results of the analyses, see Supplementary material 7.

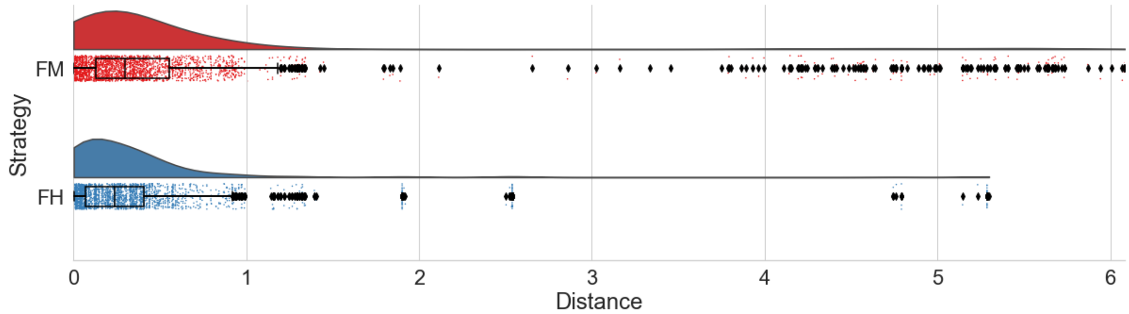


Fig. 19 The raincloud plots of distributions of the distances of familiarity which is compared with the decision threshold (i.e., "$||\mathbf{FamQ} - \mathbf{FamA}| - |\mathbf{FamQ} - \mathbf{FamB}||$" in familiarity-matching [FM]; while "$\mathbf{min}(|\boldsymbol{FamQ} - \boldsymbol{medianFamQs}|, |\boldsymbol{FamA} - \boldsymbol{FamB}|)$" in familiarity heuristic [FH]). If the distance is larger, heuristics can be used even when a person's decision threshold is higher (i.e., a lower sensitivity is required). In short, the larger (smaller) distance means that a lower (higher) sensitivity is required to use heuristics.

8.2 What is the adaptive use of heuristics? Based on the applicability for each user and the prediction from attribute substitution

So far, it was found that familiarity-matching was often used, and it would be an accurate (ecologically rational) and applicable strategy in a relationships-comparison task. However, how often were there chances to use a particular inferential strategy for each user (i.e., familiarity-matching [FM] users, familiarity heuristic [FH] users, and knowledge-based inference [KI] users) in a behavioral experiment? In Study 1 and Study 3, I identified users of strategies by model-based approaches. However, I did not focus on how often each user really had chances to use the strategy (i.e., the applicability of each strategy). In Study 4, on the other hand, I evaluated each strategy and discussed the advantage of FM in terms of applicability. However, I did not focus on how applicable the strategy was for each user (i.e., each user's applicability in a behavioral experiment). So, it is still unclear how often there were chances for each user to use a particular inferential strategy in a task set and whether or to what extent each user really used a particular strategy under the applicability.

To integrate these results, additional analyses were conducted in terms of each strategy's applicability for each user, based on Study 3's data (as in Study 4). I calculated the mean of each applicability of FM, FH, and KI, for FM, FH, and KI users in a relationships-comparison task. Note that, to discuss the use of both heuristics and knowledge, I focused on the results in difficult questions (Table 5 (A)) because only one participant was classified into FM and no participants into FH in easy questions. Also note that, statistical tests were not conducted because the numbers of classified participants were highly skewed.

Main results were as follows (Table 5 (A)). First, KI's applicability was low for all users, even for KI users; rather FM's and FH's applicability were higher than KI's applicability for all users (right column). Second, FM's applicability was generally high for all users (left column) while FH's applicability was low for FM users (middle column). The second result can corroborate the findings in behavioral experiments (i.e., FM users in difficult questions had higher decision threshold [i.e., lower sensitivity] than FH users; Table 3 (A)) and those in computer simulations (i.e., higher decision threshold would be required in using FM; Fig. 16 (A)).

Table 5 Results of additional analyses on each strategy's applicability for each user, using Study 3's data. (A) Difficult questions. (B) Easy questions.

Note: FM: familiarity-matching model. FH: familiarity heuristic model. KI: knowledge-based inference model. Statistical tests were not conducted because of highly skewed sample sizes.

(A)

Difficult	FM applicability	FH applicability	KI applicability
FM user (n = 35)	.93	.48	.27
FH user (n = 6)	.97	.97	.35
KI user (n = 6)	.94	.81	.33

(B)

Easy	FM applicability	FH applicability	KI applicability
FM user (n = 1)	.94	.94	.40
FH user (n = 0)	--	--	--
KI user (n = 50)	.90	.78	.73

Based on these results, one may wonder whether people really showed the adaptive use of heuristics. In a relationships–comparison task (Study 1 and Study 3), people often used FM in difficult questions and KI in easy questions. However, even in easy questions, FM had higher applicability than KI (see Fig. 14 and Table 5) and the correct rate of FM was as high as that of KI. So, using FM constantly (i.e., using FM regardless of the difficulty levels of questions) may be more "adaptive" than switching FM and KI (i.e., using FM in difficult questions and using KI in easy questions). Can this switching be regarded as the adaptive use?

I consider that I can explain this issue based on the attribute substitution framework (e.g., Honda et al., 2017; see also Fig. 6 in this thesis). According to the attribute substitution, people first consider whether they can solve the question based on their knowledge about it. If they do not feel difficulty for the question (i.e., they know the correct answer or have further information about the task), then they will use KI. However, if they feel difficulty for the question (i.e., they cannot use their knowledge), then they will use FM or FH. So, it is suggested that people tried to use their knowledge before trying to use heuristics, regardless of the applicability of heuristics. In difficult questions, many people could not use their knowledge (i.e., low KI's applicability), and when they try to use heuristics, they were more likely to use FM because FM was more applicable than FH (Fig. 20 (A)). In easy questions, on the other hand, many people could use their knowledge (i.e., high KI's applicability) and KI could lead correct inferences (in fact, many people knew the answer; see Fig. 21), so they directly made inferences without using heuristics (Fig. 20 (B); e.g., they will directly make an inference when people know the correct answer, even if |FamQ – FamA| is sufficiently different from |FamQ – FamB| and thus FM is applicable). Based on the attribute substitution, the results in this section imply as follows: If people have available further knowledge, then they use KI (ignoring the use of heuristics), and can often solve a task. However, if people do not have further knowledge about the task, then they rely on heuristics, especially they will use FM because using FM requires higher decision threshold (i.e., lower sensitivity) than using FH. Thus, these results will indicate the adaptive use of inferential strategies in a relationships-comparison task, under people's cognitive constraints (e.g., lack of knowledge, limited decision threshold).

Note that, apart from the above, I also conducted exploratory analyses about participants' familiarity and knowledge that may affect the subjective difficulty or the use of FM. Focusing on the "relationships" between object Q and correct/false alternatives (e.g., |FamQ – familiarity of correct alternative|, the number of accorded attributes between Q and alternatives), I plotted the distributions of familiarity of objects, or predicted the difficulty or the use of FM from familiarity and knowledge by using a linear mixed model. However, remarkable tendencies could not be observed. For these results, see Supplementary material 8.

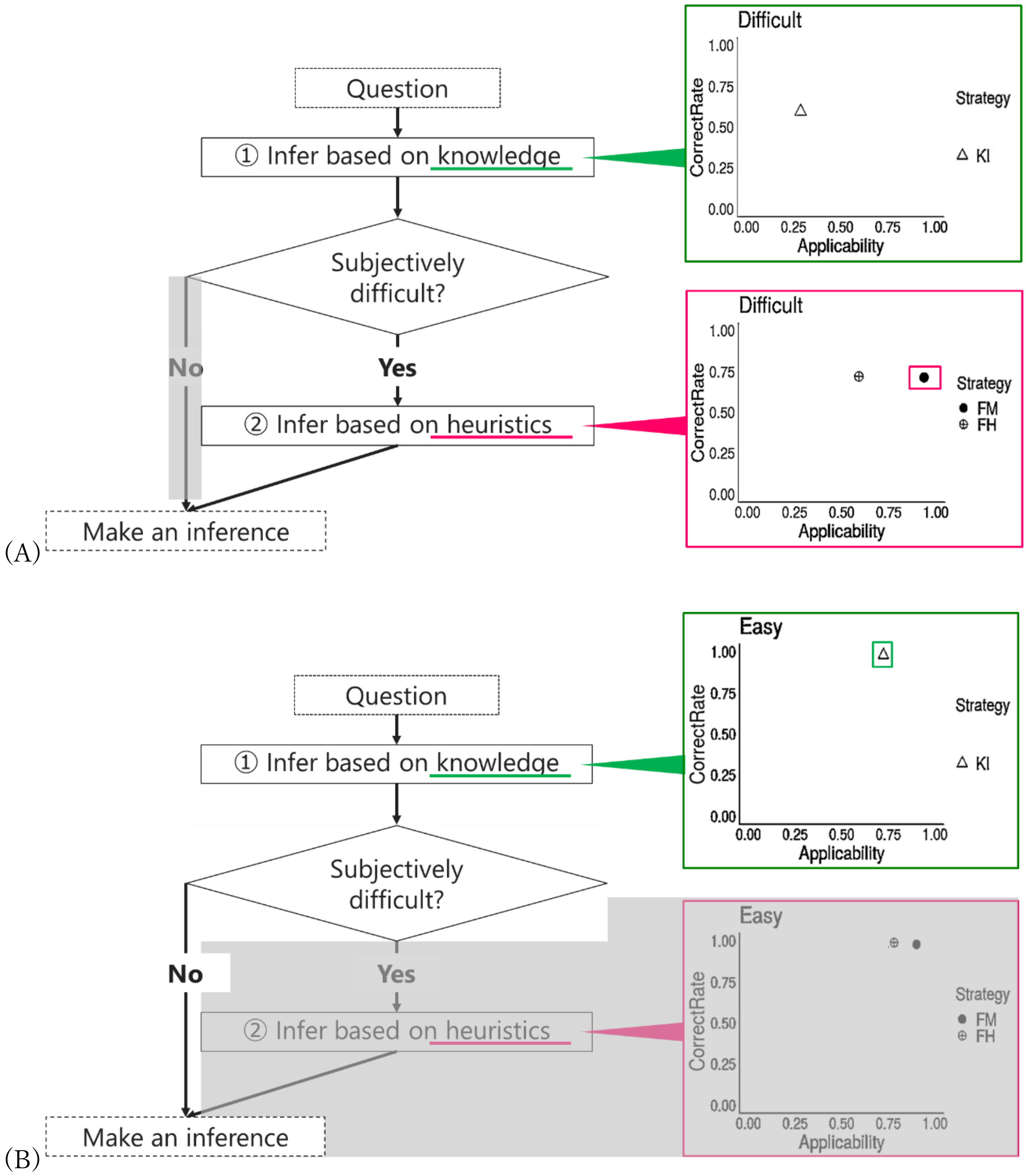


Fig. 20 Flowchart of attribute substitution framework (Fig. 6) with the applicability and correct rate of strategies (Fig. 14). According to attribute substitution, people first consider whether they can solve the question based on their knowledge about it. If they do not feel difficulty for the question, then they try to solve it using their knowledge (knowledge-based inference [KI]). However, if they feel difficulty for the question (i.e., cannot use their knowledge), then they try to solve it using heuristics (familiarity-matching [FM] or familiarity heuristic [FH]). That is, people do not consider the processes covered by the gray zones in the panels. (A) Difficult questions. Since KI cannot be applied so much (i.e., low applicability),

then people will use heuristics. And, since FM is more applicable than FH (although the correct rate of FH is as high as that of FM), then people will use FM than FH (shown in a pink square). (B) Easy questions. Since KI can be applicable well (i.e., high applicability) and will lead correct inferences (i.e., high correct rate), then people will use KI, regardless of heuristics' applicability and correct rate (shown in a green square).

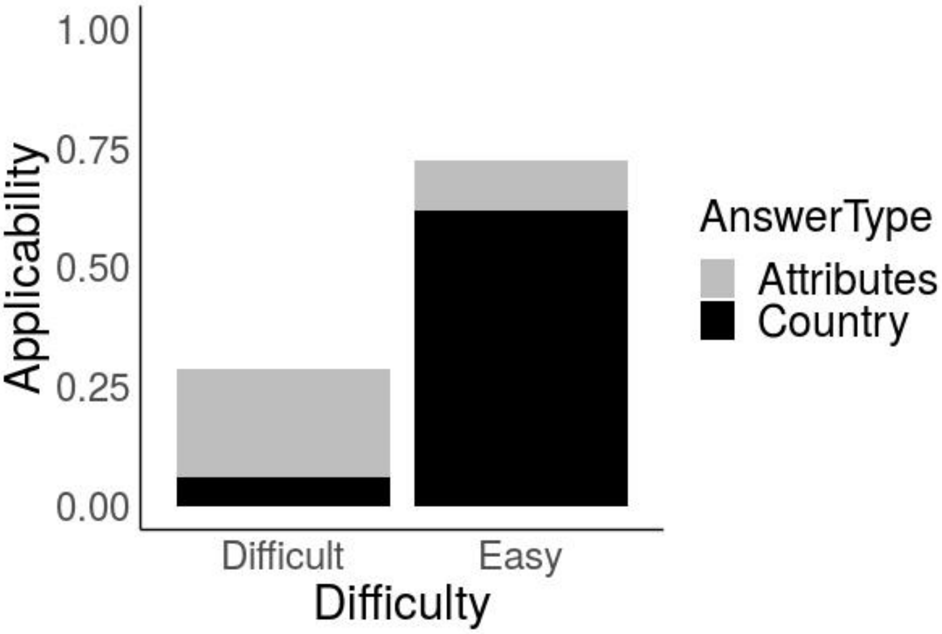


Fig. 21 The proportion of knowledge-based inference (KI) users' answers type. "Attributes" and "Country" denote the cases where participants relied on attribute cues (i.e., region, language, and religion) and the cases where they directly answered the country, respectively. Based on the attribute substitution framework (Honda et al., 2017), it was assumed that participants first considered the country, and if they did not know it, then they would rely on heuristics or attribute cues.

8.3 Related classical theories about the three-objects relationships and the "matching" behaviors

As to my relationships-comparison task and familiarity-matching (FM), it can be natural that people try to solve it considering two dyadic relationships (i.e., relating Q with A, and relating Q with B) when they look at three objects in a binary choice task. I will discuss two possible associations between my study and classical theories about "relationships" among three objects. The first issue is the structural mapping theory in analogical thinking. Gentner (1983) have discussed that there are two types of "relations" between objects: First-order relation, which uses objects as an argument; and higher-order relation, which takes relations as arguments (Gentner, 1983). In using FM for a relationships-comparison task is assumed to involve two processes: The process of calculating similarities in familiarity (i.e., computing |FamQ – FamA| and |FamQ – FamB|); and the process of comparing these calculated similarities (i.e., comparing |FamQ – FamA| with |FamQ – FamB|). These processes may correspond to higher-order relation and first-order relation in the structural mapping theory, respectively. The second issue is human cognitive developments (as review, Cook, 2018). People, even children, generally perceive their external world exploiting relationships between three objects. Furthermore, the ability to capture these associations has been regarded as human-specific intelligence. For example, pictorial depth perception about three objects based on relative heights or sizes (e.g., Biederman, Mezzanotte, & Rabinowitz, 1982), musical perception of positive or negative affections based on three-tone combinations (e.g., Kastner & Crowder, 1990), and joint attention from a triadic interaction (self, other, and object) (e.g., Saxe, 2006). Because these theories have been proposed, it can be natural that people consider the third object (e.g., city Q) for solving a relationships-comparison task (wherein three objects are presented).

In addition, people's judgments based on a "matching" behavior have also been widely observed in classical cognitive science studies. For example, as the names imply, the *matching bias* in a deduction reasoning (e.g., Evans, 1972; Evans & Lynch, 1973), or the *probability matching* in a probability judgment (e.g., Gaissmaier & Schooler, 2008; Vulkan, 2000; West & Stanovich, 2003). Another example, which seemed to be more similar to this study, is the *context effects* in similarity judgments (e.g., Medin & Schaffer, 1978; Tversky, 1977; Tversky & Gati, 1978). In previous studies, when one target and multiple alternatives were presented, people matched (grouped) the target with a different alternative if a certain alternative object changed. In particular, when Austria (target), Sweden, Poland, and Hungary (alternatives) were presented, many people thought that "Sweden" was the most similar to Austria, assuming the Cold War era between East and West when this experiment was conducted. However, if Poland has been changed to Norway, many people thought that "Hungary" was the most similar to Austria, assuming the stronger association of these countries based on historical or graphical background. Although I proposed familiarity-matching as a new heuristic in this thesis, I believe that people generally show "matching" behaviors as many previous studies reported, and that, therefore, my study is not just a new study but also a study which will be closely related to classical theories.

8.4 Limitations

I will point out four limitations of this thesis[xviii]. The first limitation is about the generality of the proposed experimental task. In Studies 1~4, I used experimental materials with a clear inclusion relationship (i.e., city and country) and focused on human inferences from memory[xix] (i.e., people have to make inferences based solely on knowledge retrieved from memory; see also Gigerenzer & Goldstein, 1996). It is still unclear to what extent the current findings can be generalized. I believe, however, that the findings can be generalized regardless of a clear or unclear inclusion relationship if a task structure (and an environmental structure that people may exploit) is identical to that in my experiments (i.e., like Fig. 1 (D)). This is because, as Gigerenzer et al. (1999) argued, heuristics are so simple that they are likely to be applied to new situations (i.e., rarely overfitting a particular situation). People make inferences adaptively depending on environments in which they are embedded (e.g., Gigerenzer, 2008; Goldstein & Gigerenzer, 2002; Todd & Gigerenzer, 2007). In fact, although the experimental materials of Study 5 in this thesis did not have a clear inclusion relationship and were focused on preferential choices, the results suggested that FM tended to be used. So, what is important will be a task structure and an environmental structure, not the hierarchical relationship of materials itself or the focus of the task (e.g., inclusion relationships or inferences from memory, respectively).

The second limitation is about the level of explanation by the proposed models. In FM model, for example, I assumed that a person computed the difference of similarity in familiarity (e.g., $|\mathrm{FamQ} - \mathrm{FamA}| < |\mathrm{FamQ} - \mathrm{FamB}|$). One may argue, however, that this assumption was inappropriate because it would be almost impossible that a person strictly computed similarity in familiarity in such a way. There might be participants who rated difficulty and familiarity sparsely, not continuously (e.g., only 2-points rating such as "0" or "100"), even though they were asked to rate them in a visual analog scale. So, was it really appropriate to apply such models to all participants? However, as described in Chapter 2, the type of the models in this thesis is "as-if" model, and the models provide explanations at a computational level. So, I only estimated individuals' inferential strategy, and did not assume that participants made inferences in the exact same way as the model did. To the best of my knowledge, no previous studies have reported that the performances of model fitting will drastically decrease due to such "sparse" ratings. However, to address the above argument, I conducted additional analyses using Study 1's data. As a result, it is implied that many participants did not rate difficulty and familiarity sparsely, and that the values of the goodness of fit, G^2, did not so much differ between "sparse" raters and "non-sparse" raters in heuristic models (for details of the results, see Supplementary material 9). So, I believe that it was not a serious problem to apply the current heuristic models to all participants.

The third limitation is about the manipulation of familiarity. In some classical studies on human inferences, familiarity with objects was manipulated: For example, by changing the frequency of presenting objects between participants (e.g., Song & Schwarz, 2008; Whittlesea, 1993); by presenting either familiar or unfamiliar item as an experimental stimulus (e.g., Alter & Oppenheimer, 2008); or by setting the various combinations of cue values in two alternatives and asking participants to learn these cue validities (e.g., Bröder, 2003). I believe that there are merits and demerits in manipulating familiarity. One of the merits is that researchers can control participants' knowledge and subjective memory experiences. In contrast, one of the demerits is that it will be difficult for researchers to discuss ecological rationality of heuristics because researchers cannot directly focus on participants' knowledge or memory obtained though media in the real world if familiarity is experimentally manipulated. In this thesis, one of my interests was the ecological rationality according to the adaptive toolbox framework. So, I adopted the experimental procedure wherein researchers did not experimentally manipulate participants' familiarity. Based on the above merit, however, to obtain further understandings of human inferences, the procedure of experimental manipulation of familiarity may be needed in the future.

Finally, the fourth limitation is about the effect of the task order in a behavioral experiment. This issue has been discussed in many previous studies within the adaptive toolbox framework (e.g., Goldstein & Gigerenzer, 2002; Hogarth & Einhorn, 1992; Pohl, 2006). However, to the best of my knowledge, there are no clear evidence that the order of tasks can strongly affect the experimental results in this framework. I believe, therefore, that the main results in this thesis will not differ even if the task order (i.e., at first "relationships-comparison task," then "the measurement of familiarity task," and finally "knowledge task") changes. Note that, however, it is widely observed that a task or a stimulus can affect the subsequence task (e.g., anchoring effect e.g., Tversky & Kahneman, 1974). For example, it may be possible that participants feel more familiar with a certain city in the measurement of familiarity task because they saw and chose the city in the previous binary choice task. So, future work will be needed to consider the order effect of experimental tasks.

8.5 Computer simulations for further examinations of effects of human cognitive constraints

As to computer simulations using ACT-R (see Study 4b), at least two points should be needed to examine, in terms of human cognitive constraints. The first point is "forgetting (or, *decay*)." In the current simulations, I manipulated only "decision threshold (or, *sensitivity*)" parameter and fixed decay parameter, d, as a typical value (–0.5; e.g., Schooler & Hertwig, 2005). For further understandings of heuristics' adaptive nature, I should consider the effects of decay on heuristics' performances. This is because "less-is-more effect," which is a well-known phenomenon in human inferences (e.g., Goldstein & Gigerenzer, 1999; 2002), can be investigated by manipulating the decay parameter (e.g., Fechner et al., 2019; Schooler & Hertwig, 2005). Less-is-more effect explains how people will make more accurate inferences in situations where they have less knowledge or information than those where they have more. Through behavioral experiments and computer simulations, previous studies have demonstrated that people who have "moderate" knowledge tend to show the better inferential performances than those who have no or little knowledge or those who have much knowledge (e.g., Schooler & Hertwig, 2005) in classical binary choice tasks such as population inference task. However, the relationships-comparison task has a new task structure that have never been investigated in previous studies. So, in using FM for relationships-comparison tasks, it is still unknown whether less-is-more effect will be observed as in previous studies.

The second point for further examinations is to simulate knowledge-based inferences. Study 4b simulated only familiarity-based heuristics (i.e., FM and FH). However, as described in previous chapters, people sometimes make inferences using their knowledge, not familiarity, about a task. So, I should simulate human inferences considering the retrieving of their knowledge (i.e., assuming that particular knowledge is available or unavailable), and should evaluate effects of the amount of available knowledge on inferential performances. In fact, Marewski and Schooler (2011) constructed a model for probability of retrieving knowledge about an object in general knowledge tasks and examined heuristics' performances (see also Dimov & Marewski, 2017; Marewski & Mehlhorn, 2011). For simulating knowledge-based inferences, I believe that these models may be able to be applied.

By simulating how the decay and knowledge affect people's adaptive use of heuristics in a relationships-comparison task, I may reveal the adaptive nature of human cognitive constraints in a new task structure.

Chapter 9 Conclusion

According to the framework of *adaptive toolbox*, a simple heuristic can be a useful inferential strategy when the structure of the heuristic (i.e., "cognition" blade in Simon's scissors) and that of an environment (i.e., "context" blade in Simon's scissors) fit together well and when the heuristic is more applicable than other strategies under people's cognitive constraints. In this thesis, I focused on a *task structure* of a binary choice as a new context aspect. I first predicted as follows: If a task structure differed from a structure in previous studies, then an environmental structure that people might exploit would differ, and an adaptive heuristic would also differ. To obtain further understandings of heuristics within the adaptive toolbox framework, I proposed a binary choice task that has a new task structure (i.e., wherein not only two objects in alternatives but also an object in a question were presented), *relationships-comparison task*. In this thesis, three inferential models were constructed to describe human inferences for a relationships-comparison task (*familiarity-matching* [FM], *familiarity heuristic* [FH], and *knowledge-based inference* [KI]). Using a relationships-comparison task and model-based approaches, I examined the strategy that people use, the accuracy and the applicability of these strategies. The main results of these three focuses are as follows:

The strategy that people use (estimated from the behavioral experiments): In Study 1, it was found that people were more likely to make inferences based on a similarity in familiarity (i.e., FM) in a relationships-comparison task. The main findings of Study 1 could be replicated in Study 3. Furthermore, in Study 5, I obtained the first evidence that people tended to use FM even in a daily context (i.e., consumer choices).

The accuracy (clarified from the behavioral experiment, data analyses, and computer simulations): In Study 2, the results showed that familiarity-matching was an ecologically rational heuristic. That is, the structure of FM could match to that of the real-world environment well, and FM could lead many correct inferences.

The applicability (clarified from the data analyses and computer simulations): In

Study 4, I revealed that FM could be more applicable than the other strategies in a relationships-comparison task, regardless of people's decision threshold (or, sensitivity) for discriminating between similarities in familiarity. That is, using FM required a higher decision threshold (or, lower sensitivity).

It may be natural that a new heuristic was used when a task structure changed (i.e., relationships-comparison task). However, it is noteworthy that this new heuristic had not only high accuracy but also high applicability. According to the adaptive toolbox, these results mean that people intuitively "selected" a more accurate and applicable strategy among several strategies, even in a completely new task structure. I believe that I could provide further understandings of people's adaptive use of heuristics because I clarified the adaptive toolbox framework held true in a new task structure. As a take home message, because human intelligence has the ability to capture relationships between three objects (see also section 8.3), it is important for researchers to pay more attention to a task structure as a "context" blade of Simon's scissors in order to understand the adaptive use of heuristics in the adaptive toolbox framework.

Acknowledgement

All data and code in this thesis are available on the Open Science Framework https://osf.io/5cmfe/?view_only=a099c9826d874d71a84f4414caf66f67.

I would like to thank everyone who supported and guided me during writing my thesis. First, I would like to thank my supervisor, Kazuhiro Ueda. I have been supported by him for five years including my time as a master's student. Although he was always very busy, he gave me many constructive advices whenever I needed the help. I really learned a lot from him.

I also thank the reviewers of my thesis, Yurika Yokoyama, Kazuo Hiraki, Sachiko Kiyokawa, and Kuninori Nakamura for their many constructive comments in my thesis defense and my thesis explanations prior to it.

I would also like to thank my collaborators, Toshihiko Matsuka, and Hidehito Honda. They have supported me since I was an undergraduate student, and also have provided helpful insights about experimental designs, cognitive modeling, data analyses, and so on.

Moreover, I would like to thank an action editor of our academic journal paper (Shirasuna et al., 2020, *Cognitive Science*). Thanks to the constructive advice, I think that I could significantly improve our academic paper and could write my thesis based on the paper.

Many thanks to all members of Ueda laboratory, too. Especially, Yuri Sato (researcher), Shuma Iwatani (researcher), Itsuki Fujisaki (graduate student), Kunhao Yang (graduate student), Taku Imaizumi (graduate student), Arata Nonami (graduate student), Ryotaro Sasaki (undergraduate student), and Qixiang Chen (alumnus). They provided me not only many comments about my studies but also many fun hours inside and outside the laboratory.

Apart from the academia, I also thank my friends and fellows, especially, Shingo Tsuemura, Masahiko Tanaka, Hiroki Moriya, and Moe Yokokura. They often supported me in many situations: Participating my pilot experiments for many times, providing advice about my oral presentation, discussing my studies, helping my office work, going to eat, and simply talking with me.

Finally, but not least, I am grateful to my family and relatives. Although they all live in faraway places from me (and although recently I have not been able to see them, even my parents and my sister, due to the corona virus pandemic), they all encouraged me in my difficult times and shared happiness in good times. Especially, I am very grateful to my twin brother, Hitoshi Shirasuna, for living with me and supporting me in my everyday life (often, participating in my pilot experiments). Thank you so much.

References

Alter, A. L., & Oppenheimer, D. M. (2008). Easy on the mind, easy on the wallet: The roles of familiarity and processing fluency in valuation judgments. *Psychonomic Bulletin & Review*, *15*(5), 985–990. https://doi.org/10.3758/PBR.15.5.985

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060. https://doi.org/10.1037/0033-295X.111.4.1036

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177. https://doi.org/10.1016/0010-0285(82)90007-X

Binz, M., Gershman, S. J., Schulz, E., & Endres, D. (2020). Heuristics From Bounded Meta-Learned Inference. *PsyArXiv*, 1–63. https://doi.org/10.1017/CBO9781107415324.004

Blokpoel, M., Kwisthout, J., van der Weide, T. P., Wareham, T., & van Rooij, I. (2013). A computational-level explanation of the speed of goal inference. *Journal of Mathematical Psychology*, *57*(3–4), 117–133. https://doi.org/10.1016/j.jmp.2013.05.006

Brighton, H. (2020). Statistical Foundations of Ecological Rationality. *Economics: The Open-Access, Open-Assessment E-Journal*, *14*(2020–2), 1–32. Retrieved from http://www.economics-ejournal.org/economics/journalarticles/2020-2

Bröder, A. (2003). Decision Making with the “Adaptive Toolbox”: Influence of Environmental Structure, Intelligence, and Working Memory Load. *Journal of Experimental Psychology: Learning Memory and Cognition*, *29*(4), 611–625. https://doi.org/10.1037/0278-7393.29.4.611

Bröder, A., & Schiffer, S. (2003). Bayesian strategy assessment in multi-attribute decision making. *Journal of Behavioral Decision Making*, *16*(3), 193–213. https://doi.org/10.1002/bdm.442

Castela, M., Kellen, D., Erdfelder, E., & Hilbig, B. E. (2014). The impact of subjective recognition experiences on recognition heuristic use: A multinomial processing tree approach. *Psychonomic Bulletin and Review*, *21*(5), 1131–1138.

https://doi.org/10.3758/s13423-014-0587-4

Chase, V. M., Hertwig, R., & Gigerenzer, G. (1998). Visions of rationality. *Trends in Cognitive Sciences*, *2*(6), 206–214.

Chater, N., Felin, T., Funder, D. C., Gigerenzer, G., Koenderink, J. J., Krueger, J. I., … Todd, P. M. (2018). Mind, rationality, and cognition: An interdisciplinary debate. *Psychonomic Bulletin & Review*, *25*(2), 793–826. https://doi.org/10.3758/s13423-017-1333-5

Clark, G. L. (2010). Human nature, the environment, and behaviour: explaining the scope and geographical scale of financial decision - making. *Geografiska Annaler: Series B, Human Geography*, *92*(2), 159–173. https://doi.org/10.1111/j.1468-0467.2010.00340.x

Cook, N. D. (2018). The Triadic Roots of Human Cognition: "Mind" Is the Ability to go Beyond Dyadic Associations. *Frontiers in Psychology*, *9*, 1060. https://doi.org/10.3389/fpsyg.2018.01060

Cornil, Y., Hardisty, D. J., & Bart, Y. (2019). Easy, breezy, risky: Lay investors fail to diversify because correlated assets feel more fluent and less risky. *Organizational Behavior and Human Decision Processes*, *153*(June), 103–117. https://doi.org/10.1016/j.obhdp.2019.06.001

Curley, L. J., MacLean, R., Murray, J., & Laybourn, P. (2019). Decision Science: A New Hope. *Psychological Reports*, *122*(6), 2417–2439. https://doi.org/10.1177/0033294118797579

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In P. In G. Gigerenzer & & the A. R. G. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 97–118). New York: Oxford University Press. Retrieved from http://www.dangoldstein.com/papers/Czerlinski_How_Good_Are_Simple_Heuristics_Chapter_1999.pdf

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. Retrieved from http://web.b.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=0&sid=e34995c6-4b18-4a1f-99eb-6e434e393c16%40sessionmgr120

del Campo, C., Pauser, S., Steiner, E., & Vetschera, R. (2016). Decision making styles and the use of heuristics in decision making. *Journal of Business Economics*, *86*(4), 389–412. https://doi.org/10.1007/s11573-016-0811-y

Dimov, C. M., & Marewski, J. N. (2017). Architectural process models of decision making: Towards a model database ACT-R. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1931–1936). Retrieved from https://serval.unil.ch/resource/serval:BIB_A19A74C97B15.P001/REF.pdf

Durbach, I. N., Algorta, S., Kabongo, D., Katsikopoulos, K. V, & Şimşek, Ö. (2020). Fast and frugal heuristics for portfolio decisions with positive project interactions. *Decision Support*

Systems, *138*(April), 113399. https://doi.org/10.1016/j.dss.2020.113399
Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial Processing Tree Models A Review of the Literature. *Zeitschrift Für Psychologie/Journal of Psychology*, *217*(3), 108–124. https://doi.org/10.1027/0044-3409.217.3.108
Evans, J. (1972). Interpretation and Matching Bias in a Selectioning Task. *Quarterly Journal of Experimental Psychology*, *24*, 193–199. Retrieved from http://journals.sagepub.com/doi/pdf/10.1080/00335557243000067
Evans, J. B. S., & Lynch, J. S. (1973). Matching Bias in the Selection Task. *British Journal of Psychology*, *64*(3), 391–397. https://doi.org/10.1111/j.2044-8295.1973.tb01365.x
Fechner, H. B., Pachur, T., & Schooler, L. J. (2019). How does aging impact decision making? The contribution of cognitive decline and strategic compensation revealed in a cognitive architecture. *Journal of Experimental Psychology: Learning Memory and Cognition*, *45*(9), 1634–1663. https://doi.org/10.1037/xlm0000661
Fechner, H. B., Schooler, L. J., & Pachur, T. (2018). Cognitive costs of decision-making strategies: A resource demand decomposition analysis with a cognitive architecture. *Cognition*, *170*, 102–122. https://doi.org/10.1016/j.cognition.2017.09.003
Forbes, W., Hudson, R., Skerratt, L., & Soufian, M. (2015). Which heuristics can aid financial-decision-making? *International Review of Financial Analysis*, *42*, 199–210. https://doi.org/10.1016/j.irfa.2015.07.002
Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*(3), 416–422. https://doi.org/10.1016/j.cognition.2008.09.007
Gentner, D. (1983). Structure Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, *7*(2), 155–170.
Ghezzi, A. (2020). How Entrepreneurs make sense of Lean Startup Approaches: Business Models as cognitive lenses to generate fast and frugal Heuristics. *Technological Forecasting and Social Change*, *161*(September), 120324. https://doi.org/10.1016/j.techfore.2020.120324
Gigerenzer, G. (2001). Bounded Rationality - The Adaptive Toolbox: Chapter 3. In G. Gigerenzer & R. Selten (Eds.), *Bounded Rationality - The Adaptive Toolbox* (pp. 36–50). The MIT Press. Retrieved from http://pubman.mpdl.mpg.de/pubman/item/escidoc:2102373/component/escidoc:2102372/GG_Adaptive_Toolbox_2001.pdf
Gigerenzer, G. (2008). Why Heuristics Work. *Psychological Science*, *3*(1), 20–29. Retrieved from http://journals.sagepub.com/doi/pdf/10.1111/j.1745-6916.2008.00058.x
Gigerenzer, G. (2019). How to Explain Behavior? *Topics in Cognitive Science*, 1–19.

https://doi.org/10.1111/tops.12480

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, *1*(1), 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology* , *62*, 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the Fast and Frugal Way: Models of Bounded Rationality. *Psychological Review*, *103*(4), 650–669. https://doi.org/10.1093/acprof

Gigerenzer, G., & Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making*, *6*(1), 100–121. Retrieved from http://library.mpib-berlin.mpg.de/ft/gg/GG_Recognition_2011.pdf

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press. Retrieved from http://pubman.mpdl.mpg.de/pubman/item/escidoc:2102905/component/escidoc:2102904/GG_Fast_1999.pdf

Gigerenzer, G., Todd, P. M., & ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA: Oxford University Press, USA.

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In *Simple heuristics that make us smart* (pp. 37–58). New York: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of Ecological Rationality: The Recognition Heuristic. *Psychological Review*, *109*(1), 75–90. https://doi.org/10.1037//0033-295X.109.1.75

Griffiths, T. L. (2020). Understanding Human Intelligence through Human Limitations. *PsyArXiv*, (3), 55. Retrieved from https://arxiv.org/pdf/2009.14050

Hauser, J. R. (2014). Consideration-set heuristics. *Journal of Business Research*, *67*(8), 1688–1699. https://doi.org/10.1016/J.JBUSRES.2014.02.015

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1191–1206. https://doi.org/10.1037/a0013025

Hertwig, R., Pachur, T., & Kurzenhäuser, S. (2005). Judgments of Risk Frequencies: Tests of Possible Cognitive Mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 621–642. https://doi.org/10.1037/0278-7393.31.4.621

Hilbig, B. E. (2014). On the role of recognition in consumer choice: A model comparison.

Judgment and Decision Making, *9*(1), 51–57. Retrieved from https://ideas.repec.org/a/jdm/journl/v9y2014i1p51-57.html

Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2011). Fluent, fast, and frugal? A formal model evaluation of the interplay between memory, fluency, and comparative judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 827–839. https://doi.org/10.1037/a0022638

Hilbig, B. E., Michalkiewicz, M., Castela, M., Pohl, R. F., & Erdfelder, E. (2015). Whatever the cost? Information integration in memory-based inferences depends on cognitive effort. *Memory & Cognition*, *43*(4), 659–671. https://doi.org/10.3758/s13421-014-0493-z

Hilbig, B. E., & Pohl, R. F. (2008). Recognizing users of the recognition heuristic. *Experimental Psychology*, *55*(6), 394–401. https://doi.org/10.1027/1618-3169.55.6.394

Hilbig, B. E., & Pohl, R. F. (2009). Ignorance- versus evidence-based decision making: A decision time analysis of the recognition heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1296–1305. https://doi.org/10.1037/a0016565

Hoffmann, J. A., Bettina, von H., & Rieskamp, J. (2019). Testing learning mechanisms of rule-based judgment. *Decision*, *6*(4), 305–334.

Hoffrage, U., & Reimer, T. (2004a). Models of Bounded Rationality: The Approach of Fast and Frugal Heuristics. *Management Revue*, *15*(4), 437–459. https://doi.org/10.2307/41783487

Hoffrage, U., & Reimer, T. (2004b). Models of Bounded Rationality: The Approach of Fast and Frugal Heuristics. *Management Revu*, *15*(4), 437–459. https://doi.org/10.5771/0935-9915-2004-4-437

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55. https://doi.org/10.1016/0010-0285(92)90002-J

Honda, H., Abe, K., Matsuka, T., & Yamagishi, K. (2011). The role of familiarity in binary choice inferences. *Memory & Cognition*, *39*(5), 851–863. https://doi.org/10.3758/s13421-010-0057-9

Honda, H., Matsuka, T., & Ueda, K. (2017). Memory-Based Simple Heuristics as Attribute Substitution: Competitive Tests of Binary Choice Inference Models. *Cognitive Science*, *41*(5), 1093–1118. https://doi.org/10.1111/cogs.12395

Jenny, M. A., Rieskamp, J., & Nilsson, H. (2014). Inferring conjunctive probabilities from noisy samples: Evidence for the configural weighted average model. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(1), 203–217. https://doi.org/10.1037/a0034261

Kahneman, D., & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases* (pp. 49–81). Cambridge University Press.

https://doi.org/10.1017/cbo9780511808098.004
Kahneman, D., & Frederick, S. (2005). A Model of Heuristic Judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–293). Cambridge University Press. Retrieved from https://psycnet.apa.org/record/2005-09680-012
Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3
Kastner, M. P., & Crowder, R. G. (1990). Perception of the Major/Minor Distinction: IV. Emotional Connotations in Young Children. *Music Perception*, *8*(2), 189–201. https://doi.org/10.2307/40285496
Kozyreva, A., & Hertwig, R. (2019). The interpretation of uncertainty in ecological rationality. *Synthese*, 1–31. https://doi.org/10.1007/s11229-019-02140-w
Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and Deliberate Judgments Are Based on Common Principles. *Psychological Review*, *118*(1), 97–109. https://doi.org/10.1037/a0020762
Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building Machines That Learn and Think Like People. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X16001837
Lee, M. D., Gluck, K. A., & Walsh, M. M. (2019). Understanding the complexity of simple decisions: Modeling multiple behaviors and switching strategies. *Decision*, *6*(4), 335–368. https://doi.org/10.1037/dec0000105
Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*(e1), 1–60. https://doi.org/10.1017/S0140525X1900061X
Lockton, D. (2012). Simon's Scissors and Ecological Psychology in Design for Behaviour Change. *SSRN Electronic Journal*, 1–13. https://doi.org/10.2139/ssrn.2125405
Long, A. R., Fernbach, P. M., & De Langhe, B. (2018). Circle of incompetence: Sense of understanding as an improper guide to investment risk. *Journal of Marketing Research*, *55*(4), 474–488. https://doi.org/10.1509/jmr.16.0429
Loock, M., & Hinnen, G. (2015). Heuristics in organizations: A review and a research agenda. *Journal of Business Research*, *68*(9), 2027–2036. https://doi.org/10.1016/j.jbusres.2015.02.016
Marewski, J. N., Gaissmaier, W., Schooler, L. J., Goldstein, D. G., & Gigerenzer, G. (2009). Do Voters Use Episodic Knowledge to Rely on Recognition? Recognition in Political Elections: A Study. In *31st Annual Conference of the Cognitive Science Society* (pp. 2232–2237). https://doi.org/citeulike-article-id:8756890

Marewski, J. N., & Mehlhorn, K. (2011). Using the ACT-R architecture to specify 39 quantitative process models of decision making. *Judgment and Decision Making*, *6*(6), 439–519. Retrieved from https://pure.mpg.de/rest/items/item_2099219/component/file_2343867/content

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review*, *118*(3), 393–437. https://doi.org/10.1037/a0024143

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information. Vision: a computational investigation into the human representation and processing of visual information.* MIT press. https://doi.org/10.1016/0022-2496(83)90030-5

McCloy, R., Beaman, C. P., & Smith, P. T. (2008). The relative success of recognition-based interference in multichoice decisions. *Cognitive Science*, *32*(6), 1037–1048. https://doi.org/10.1080/03640210802152319

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Mohnert, F., Pachur, T., & Lieder, F. (2019). What' s in the Adaptive Toolbox and How Do People Choose From It? Rational Models of Strategy Selection in Risky Choice. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 2378–2384). Austin, TX: Cognitive Science Society. Retrieved from https://cogsci.mindmodeling.org/2019/papers/0412/0412.pdf

Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem solving program. In *IFIP congress* (Vol. 256, p. 64). https://doi.org/10.1038/224923a0

Newell, B. R. (2005). Re-visions of rationality? *Trends in Cognitive Sciences*, *9*(1), 11–15. https://doi.org/10.1016/j.tics.2004.11.005

Newell, B. R., & Bröder, A. (2008). Cognitive processes , models and metaphors in decision research. *Judgment and Decision Making*, *3*(3), 195–204. Retrieved from http://www.sas.upenn.edu/~baron/journal/bn1.pdf

Nickles, T. (2018). Bounded rationality, scissors, crowbars, and pragmatism: reflections on Herbert Simon. *Mind & Society*, *17*(1–2), 85–96. https://doi.org/10.1007/s11299-019-00206-3

Otworowska, M., Blokpoel, M., Sweers, M., Wareham, T., & van Rooij, I. (2018). Demons of Ecological Rationality. *Cognitive Science*, *42*(3), 1057–1066. https://doi.org/10.1111/cogs.12530

Pachur, T., & Aebi-Forrer, E. (2013). Selection of Decision Strategies After Conscious and Unconscious Thought. *The Journal of Behavioral Decision Making*, *26*(5), 477–488. https://doi.org/10.1002/bdm

Pachur, T., & Hertwig, R. (2006). On the psychology of the recognition heuristic: Retrieval primacy as a key determinant of its use. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 983–1002. Retrieved from http://web.a.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=0&sid=fa27680e-81ab-4345-8b9e-72a8d51fb9b2%40sessionmgr4006

Pachur, T., Hertwig, R., Gigerenzer, G., & Brandstätter, E. (2013). Testing process predictions of models of risky choice: a quantitative model comparison approach. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00646

Pachur, T., Hertwig, R., & Rieskamp, J. (2013). Intuitive judgments of social statistics: How exhaustive does sampling need to be? *Journal of Experimental Social Psychology*, *49*, 1059–1077. https://doi.org/10.1016/j.jesp.2013.07.004

Pachur, T., Todd, P. M., Gigerenzer, G., Schooler, L. J., & Goldstein, D. G. (2011). The recognition heuristic: a review of theory and tests. *Frontiers in Psychology*, *147*(2), 1–14. https://doi.org/10.3389/fpsyg.2011.00147

Parpart, P., Jones, M., & Love, B. C. (2018). Heuristics as Bayesian inference under extreme priors. *Cognitive Psychology*, *102*, 127–144. https://doi.org/10.1016/j.cogpsych.2017.11.006

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge, England: Cambridge University Press. Retrieved from https://pdfs.semanticscholar.org/b33a/a05ad552b107ce2a062dfedb3277ad8724b3.pdf

Phillips, N., Hertwig, R., Kareev, Y., & Avrahami, J. (2014). Rivals in the dark: How competition influences search in decisions under uncertainty. *Cognition*, *133*(1), 104–119. https://doi.org/10.1016/j.cognition.2014.06.006

Pohl, R. F. (2006). Empirical tests of the recognition heuristic. *Journal of Behavioral Decision Making*, *19*(3), 251–271. https://doi.org/10.1002/bdm.522

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. Retrieved from https://www.stat.washington.edu/raftery/Research/PDF/socmeth1995.pdf

Rahnev, D. (2020). Confidence in the Real World. *Trends in Cognitive Sciences*, 1–2. https://doi.org/10.1016/j.tics.2020.05.005

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Rich, P., Blokpoel, M., Haan, R. De, & van Rooij, I. (2020). How Intractability Spans the Cognitive and Evolutionary Levels of Explanation. *Topics in Cognitive Science*, 1–24. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/tops.12506

Richter, T., & Späth, P. (2006). Recognition is used as one cue among others in judgment and decision making. *Journal of Experimental Psychology: Learning Memory and Cognition*, *32*(1), 150–162. https://doi.org/10.1037/0278-7393.32.1.150

Rieskamp, J., & Reimer, T. (2007). Ecological rationality. *Encyclopedia of Social Psychology*, 273–275. https://doi.org/10.1080/00207238308710058

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded Cognition: An Integrated Theory of Concurrent Multitasking. *Psychological Review*, *115*(1), 101–130. https://doi.org/10.1037/0033-295X.115.1.101

Sang, K., Todd, P. M., Goldstone, R. L., & Hills, T. (2020). Simple Threshold Rules Solve Explore/Exploit Trade-offs in a Resource Accumulation Search Task. *Cognitive Science*, *44*, e12817. https://doi.org/10.1111/cogs.12817

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, *16*(2), 235–239. https://doi.org/10.1016/j.conb.2006.03.001

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review*, *120*(1), 39–64. https://doi.org/10.1037/a0030777

Schneider, D. W., & Anderson, J. R. (2012). Modeling fan effects on the time course of associative recognition. *Cognitive Psychology*, *64*(3), 127–160. https://doi.org/10.1016/j.cogpsych.2011.11.001

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review*, *112*(3), 610–628. https://doi.org/10.1037/0033-295X.112.3.610

Schurz, G., & Hertwig, R. (2019). Cognitive Success: A Consequentialist Account of Rationality in Cognition. *Topics in Cognitive Science*, *11*(1), 7–36. https://doi.org/10.1111/tops.12410

Schwarz, N. (2011). Feelings-as-Information Theory. In *Handbook of theories of social psychology* (pp. 289–308). Retrieved from https://www.researchgate.net/profile/Norbert_Schwarz2/publication/266330418_Feelings-as-Information_Theory/links/54b868100cf2c27adc48b0f6/Feelings-as-Information-Theory.pdf

Schweickart, O., & Brown, N. R. (2014). Magnitude comparison extended: How lack of knowledge informs comparative judgments under uncertainty. *Journal of Experimental Psychology: General*, *143*(1), 273–294. https://doi.org/10.1037/a0031451

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, *41*(1), 1–20. https://doi.org/10.1146/annurev.biochem.64.1.721

Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist*, *26*(2), 145–159. https://doi.org/10.1037/h0030806

Song, H., & Schwarz, N. (2008). Fluency and the detection of misleading questions: low processing fluency attenuates the moses illusion. *Social Cognition*, *26*(6), 791–799. Retrieved from https://guilfordjournals.com/doi/pdfplus/10.1521/soco.2008.26.6.791

Spiliopoulos, L., & Hertwig, R. (2019). A Map of Ecologically Rational Heuristics for Uncertain Strategic Worlds. *Psychological Review*. https://doi.org/10.1037/rev0000171

Thoma, V., & Williams, A. (2013). The devil you know: The effect of brand recognition and product ratings on consumer choice. *Judgment and Decision Making*, *8*(1), 34–44. Retrieved from http://www.decisionsciencenews.com/sjdm/journal.sjdm.org/12/12703/jdm12703.pdf

Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The Computational Limits of Deep Learning. *ArXiv*. Retrieved from http://arxiv.org/abs/2007.05558

Todd, P. M., & Brighton, H. (2016). Building the Theory of Ecological Rationality. *Minds and Machines*, *26*(1–2), 9–30. https://doi.org/10.1007/s11023-015-9371-0

Todd, P. M., & Gigerenzer, G. (2000). Précis of Simple heuristics that make us smart. *Behavioral and Brain Sciences*, *23*(5), 727–780. https://doi.org/10.1017/S0140525X00003447

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, *16*(3), 167–171. https://doi.org/10.1111/j.1467-8721.2007.00497.x

Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality : intelligence in the world*. Oxford University Press. Retrieved from https://books.google.co.jp/books?hl=en&lr=&id=lO4EdShcg7AC&oi=fnd&pg=PP1&dq=info:VvCCTp0F7JMJ:scholar.google.com&ots=5mqyNEDg3c&sig=n42vR1VqsR1he8JytY1lVLbDQvw&redir_esc=y#v=onepage&q&f=false

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Tversky, A., & Gati, I. (1978). *Studies of Similarity*. (E. Rosch & B. Lloyd, Eds.), *Cognition and Categorization*.

Tversky, A., & Kahneman, D. (1973). Availability: A Heuristic for Judging Frequency and Probability. *Cognitive Psychology*, *5*(2), 207–232. Retrieved from https://ac.els-cdn.com/0010028573900339/1-s2.0-0010028573900339-main.pdf?_tid=ce7040a4-a4cb-4110-846a-7d3f8da47e08&acdnat=1522399112_1b6a29caaa831dc4fb233b6c42948183

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–31. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice.

Science, *211*(4481), 453–458.
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://doi.org/10.1007/BF00712851
van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science*, *32*(6), 939–984. https://doi.org/10.1080/03640210801897856
van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, 1–16. https://doi.org/10.31234/osf.io/7qbpr
van Rooij, I., Wright, C. D., Kwisthout, J., & Wareham, T. (2018). Rational analysis, intractability, and the prospects of 'as if'-explanations. *Synthese*, *195*(2), 491–510. https://doi.org/10.1007/s11229-014-0532-0
Vlek, C. A. J. (1970). Multiple probability learning. Associating events with their probabilities of occurrence. *Acta Psychologica*, *33*, 207–232. https://doi.org/10.1016/0001-6918(70)90134-4
Vulkan, N. (2000). An Economist's Perspective on Probability matching. *Journal of Economic Surveys*, *14*(7), 101–118. https://doi.org/10.1111/1467-6419.00106
Weber, E. U., & Johnson, E. J. (2009). Mindful Judgment and Decision Making. *Annual Review of Psychology*, *60*(1), 53–85. https://doi.org/10.1146/annurev.psych.60.110707.163633
West, R. F., & Stanovich, K. E. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory and Cognition*, *31*(2), 243–251. https://doi.org/10.3758/BF03194383
Whittlesea, B. W. A. (1993). Illusions of Familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(6), 1235–1253. Retrieved from https://pdfs.semanticscholar.org/0f39/c8ddbb55b82be9547514e4b5530312b684f5.pdf
Xu, P., González-Vallejo, C., Weinhardt, J., Chimeli, J., & Karadogan, F. (2018). Use of the familiarity difference cue in inferential judgments. *Memory & Cognition*, *46*, 298–314. https://doi.org/10.3758/s13421-017-0765-5

Supplementary materials

S.1 Supplementary material 1

The binary choice task in Study 1 was generated using the following four steps. Words in parentheses are Japanese descriptions, which were presented to participants in Study 1.

1. For "objects as alternatives," I selected 20 countries (more than two countries from five regions: Asia, Europe, Africa, North America, and South America) and randomly assigned these 20 countries to two groups: "Alternative A" and "Alternative B" (each group consisted of different 10 countries).

Alternative A	**Alternative B**
America (アメリカ)	Canada (カナダ)
Sweden (スウェーデン)	Bolivia (ボリビア)
Mexico (メキシコ)	Italia (イタリア)
Columbia (コロンビア)	Ukraine (ウクライナ)
Holland (オランダ)	Switzerland (スイス)
Egypt (エジプト)	Iran (イラン)
Turkey (トルコ)	Spain (スペイン)
Saudi Arabia (サウジアラビア)	Kazakhstan (カザフスタン)
Australia (オーストラリア)	New Zealand (ニュージーランド)
Mali (マリ)	Morocco (モロッコ)

2. For each alternative A (B), I made pairs with six countries, which were randomly selected from alternative B (A).

3. From each country (20 objects as alternatives), I selected 6 cities (total of 20 × 6 = 120 cities) using the following criteria:

(I) Out of the five cities, I selected the two with the largest population size in the country.

(II) For the remaining four cities, I selected cities that satisfied one (or more) of the following criteria: "its population size is high," "its name is included in that of a historical treaty, conference, or a similar historical event," "it has a world heritage site," or "it has hosted the Olympic or the Paralympic Games."

Alt. A	**Cities**	**Alt. B**	**Cities**
America	New York (ニューヨーク) Los Angeles (ロサンゼルス) Chicago (シカゴ) Washington D.C. (ワシントン D.C.) Portsmouth (ポーツマス) Saint Louis (セントルイス)	Italy	Rome (ローマ) Milano (ミラノ) Napoli (ナポリ) Genoa (ジェノヴァ) Verona (ヴェローナ) Siracusa (シラクサ)
Egypt	Cairo (カイロ) Alexandria (アレクサンドリア) Giza (ギザ) Suez (スエズ) Aswan (アスワン) Rosetta (ロゼッタ)	Iran	Tehran (テヘラン) Mashhad (マシュハド) Isfahan (イスファハーン) Nahavand (ニハーヴァンド) Ramsar (ラムサール) Tabriz (タブリーズ)
Australia	Sydney (シドニー) Melbourne (メルボルン) Brisbane (ブリスベン) Canberra (キャンベラ) Adelaide (アデレード) Cairns (ケアンズ)	Ukraine	Kiev (キエフ) Kharkiv (ハルキウ) Odessa (オデッサ) Donets (ドネツ) Chernobyl (チェルノブイリ) Sevastopol (セヴァストーポリ)
Holland	Rotterdam (ロッテルダム) The Hague (ハーグ) Leiden (ライデン) Maastricht (マーストリヒト) Utrecht (ユトレヒト) Delft (デルフト)	Kazakh- stan	Almaty (アルマトイ) Astana (アスタナ) Shymkent (シムケント) Aral (アラル) Baikonur (バイコヌール) Balqash (バルハシ)
Columbia	Bogotá (ボゴタ) Medellín (メデジン)	Canada	Toronto (トロント) Montreal (モントリオール)

	Cali (カリ) Florencia (フロレンシア) Barranquilla (バランキージャ) Cúcuta (ククタ)		Calgary (カルガリー) Ottawa (オタワ) Winnipeg (ウィニペグ) Charlottetown (シャーロットタウン)
Saudi Arabia	Riyadh (リヤド) Jiddah (ジッダ) Makkah (メッカ) Medina (メディナ) Dammam (ダンマーム) Taif (ターイフ)	Switzerland	Zurich (チューリヒ) Geneva (ジュネーブ) Basel (バーゼル) Bern (ベルン) Locarno (ロカルノ) Davos (ダボス)
Sweden	Stockholm (ストックホルム) Göteborg (イェーテボリ) Malmö (マルメ) Uppsala (ウプサラ) Lulea (ルレオ) Kalmar (カルマル)	Spain	Madrid (マドリード) Barcelona (バルセロナ) Valencia (バレンシア) Sevilla (セビリア) Tordesillas (トルデシリャス) Granada (グラナダ)
Turkey	Istanbul (イスタンブル) Ankara (アンカラ) Izmir (イズミル) Bursa (ブルサ) Edirne (エディルネ) Ephesus (エフェソス)	New Zealand	Auckland (オークランド) Tauranga (タウランガ) Dunedin (ダニーデン) Palmerston North (パーマストンノース) Waitangi (ワイタンギ) Napier (ネピア)
Mali	Bamako (バマコ) Sikasso (シカソ) Kalabancoro (カラバンコロ) Timbuktu (トンブクトゥ) Djenne (ジェンネ) Gao (ガオ)	Bolivia	Santa Cruz (サンタクルス) Cochabamba (コチャバンバ) El Alto (エルアルト) Sucre (スクレ) Potosi (ポトシ) Uyuni (ウユニ)
Mexico	Ecatepec (エカテペック) Puebla (プエブラ) Tijuana (ティフアナ) Villahermosa (ビヤエルモサ) Chihuahua (チワワ) Acapulco (アカプルコ)	Morocco	Casablanca (カサブランカ) Fez (フェズ) Tangier (タンジール) Rabat (ラバト) Marrakesh (マラケシュ) Meknes (メクネス)

Note: I provided criterion (II) for two reasons. First, if all alternatives consisted of top cities in terms of population, participants would be more likely to know the answer (i.e., they would use knowledge-based inferences instead of heuristics), because it seemed that larger cities were generally famous. Second, I wanted to select cities that most participants might know by name but would not know the countries of, because participants would be likely to rely on heuristics instead of knowledge in such cases. However, even if a city satisfied criterion (I) or (II), I excluded cities whose names included the name of the country (e.g., Mexico City) or that exist in several countries (e.g., there is a Melbourne not only in Australia but also in America).

4. To make one of the two alternatives (from step 2) a correct answer, I placed one city (from step 3) into "city Q" in each problem statement "Which country is city Q is in, A or B?" (e.g., "Which country is Sikasso in, Mali or Switzerland?")

S.2 Supplementary material 2

Experimental data of Study 1 (N = 90).

Distributions for individual data (N = 90)

Correct rates

	Min.	Mean	Median	SD	Max
Correct rate in the binary choice task	.53	.73	.73	0.09	.95
Correct rate in the knowledge task (i.e., "know the correct country" cases)	.07	.42	.37	0.24	1.0

Data from the measurement of familiarity task

	Min.	Mean	Median	SD	Max
Mean FamQ (120 city names)	1.40	31.8	30.2	14.0	67.8
Mean FamA (10 country names)	2.94	74.3	76.9	19.2	99.9
Mean FamB (10 country names)	4.76	74.4	76.2	19.3	100.0

Data from the knowledge task (except for the cases where participants answered “I do not know”)

	Min.	Mean	Median	SD	Max
The number of questions answered any country	8.00	60.0	44.0	28.9	120.0
The number of questions answered any region of city	3.00	57.1	50.5	29.8	120.0
The number of questions answered any language of city	2.00	51.4	45.0	29.2	120/0
The number of questions answered any religion of city	1.00	50.3	47.5	31.6	120.0
The number of questions answered any region of alt A	4.00	114.2	120.0	14.7	120.0
The number of questions answered any language of alt A	4.00	103.5	110.0	20.3	120.0
The number of questions answered any religion of alt A	4.00	99.4	113.0	28.0	120.0
The number of questions answered any region of alt B	4.00	114.5	120.0	14.5	120.0
The number of questions answered any language of alt B	4.00	103.9	109.5	20.0	120.0
The number of questions answered any religion of alt B	4.00	99.5	112.5	27.3	120.0

Distributions for each question (total 120 questions)

Distribution of mean difficulty ratings (total 120 questions)

	Min.	Mean	Median	SD	Max
Mean difficulty ratings in total 120 questions	4.49	56.9	62.2	23.2	84.9
Mean difficulty ratings in difficult 60 questions	61.9	75.7	77.6	6.05	84.9
Mean difficulty ratings in easy 60 questions	4.49	38.1	42.0	18.1	62.5

Distribution of the number of participants ($N = 90$) who gave the correct answer for a certain question

(e.g., eighty-nine participants correctly answered the question, "which country is Washington D.C. in, America or Italy?", which was the question most participants could solve)

	Min.	Mean	Median	SD	Max
The number of participants	2.00	33.3	27.0	25.1	89.0

Distributions of how many knowledge users' choices each LEX model (3! = six cue order patterns) could predict (I show the number of participants whose value of G^2 for the LEX model was the lowest; because some participants had two or more lowest G^2 values, the sum of the numbers of participants is not always equal to the number of total knowledge users)

	Difficult questions (17 knowledge users)	Easy questions (82 knowledge users)
Region – Language – Religion	5	28
Region – Religion – Language	6	12
Language – Region – Religion	5	23
Language – Religion – Region	6	23
Religion – Region – Language	8	41
Religion – Language – Region	6	36

S.3 Supplementary material 3

The result of response time analysis in Study 1

To investigate participants' implicit cognitive processes, I conducted an exploratory analysis of response time, using a mixed linear model, for each difficulty level (see tables on the next page; DV: response time; Fixed effects: choice pattern [familiar choice or unfamiliar choice], user [heuristic or knowledge], and the order of questions; Random effect: participants who were classified into a certain inference model). Based on previous findings (e.g., Hilbig & Pohl, 2009; Pachur & Hertwig, 2006), I focused on the differences in response time between cases in which people chose the more familiar or the less familiar alternative (I call this the *familiar-choice* or *unfamiliar-choice,* respectively). In this analysis, I excluded the cases in which the difference between FamA and FamB was below her/his best decision threshold. I used the best threshold of FM or FH for FM and FH users, respectively, and the mean of the threshold of FM and FH for knowledge users. Furthermore, I also considered differences between participants who used heuristics (FM or FH) and those who used knowledge. In the end, 55 and 65 participants were analyzed for difficult and easy questions, respectively.

Note that the response time tended to become shorter as the questions proceeded (the order effect; e.g., Schweickart & Brown, 2014). Although I observed this tendency in difficult questions (the order of questions: $p < .05$) as in previous findings, I did not observe it in easy questions (the order of questions: $p = .43$). That might be because, in many easy questions, participants knew the correct answer and could make an inference quickly regardless of when the question were presented. The main effect of the choice patterns or users and the all interaction effects were not significant.

Fixed effect in difficult questions (n = 55)

	Estimate	Std. Error	Df	*t*-value	*p*-value
(intercept)	3.698e+00	3.942e-01	7.167e+02	9.38	< .05
Pattern	-2.560e-01	4.314e-01	1.007e+03	-0.593	0.5531
User	2.843e-01	7.833e-01	6.367e+02	0.363	0.7168
Order of q	-1.225e-02	5.051e-03	1.001e+03	-2.425	< .05
Pattern * user	-2.767e-01	9.256e-01	1.010e+03	-0.299	0.7650
Pattern * order	3.448e-03	5.926e-03	9.995e+02	0.582	0.5608
User * order	-4.535e-03	1.048e-02	1.001e+03	-0.433	0.6652
Pat*user*ord	2.157e-03	1.318e-02	1.003e+03	0.164	0.8700

Fixed effect in easy questions (n = 65)

	Estimate	Std. Error	Df	*t*-value	*p*-value
(intercept)	2.383e+00	7.453e-01	1.161e+02	3.198	< .05
Pattern	4.468e-02	6.214e-01	1.605e+03	0.072	0.94270
User	7.468e-01	7.609e-01	1.181e+02	0.981	0.32839
Order of q	-5.559e-03	7.012e-03	1.605e+03	-0.793	0.42798
Pattern * user	-1.801e-01	6.441e-01	1.607e+03	-0.280	0.77977
Pattern * order	8.620e-04	9.116e-03	1.605e+03	0.095	0.92468
User * order	-2.962e-03	7.177e-03	1.606e+03	-0.413	0.67985
Pat*user*ord	4.513e-03	9.439e-03	1.607e+03	0.478	0.63264

S.4 Supplementary material 4

The binary choice task in Study 3 was generated using the following four steps. Words in parentheses are Japanese descriptions, which were presented to participants in Study 3.

1. For the materials for Study 2 "the 50 countries with the highest populations in the world and their 50 capitals," I selected 25 countries (the 1st, 3rd, …, 49th country with highest population) as objects of alternatives.

2. From each country, four cities were picked up: one was the capital of the country, and the others were the three cities with the highest population in the country (a total of 125 objects were picked).

Countries	**Cities**	
China (中国)	Beijing(ペキン)	Shanghai(シャンハイ)
	Guangzhou(コウシュウ)	Tianjin(テンシン)
America (アメリカ)	Washington D.C.(ワシントン D.C.)	Los Angeles(ロサンゼルス)
	New York(ニューヨーク)	Chicago(シカゴ)
Brazil (ブラジル)	Brasilia(ブラジリア)	Rio de Janeiro(リオデジャネイロ)
	Sao Paulo(サンパウロ)	Salvador(サルヴァドール)
Nigeria (ナイジェリア)	Abuja(アブジャ)	Ibadan(イバダン)
	Lagos(ラゴス)	Kano(カノ)
Russia (ロシア)	Moscow(モスクワ)	Novosibirsk(ノヴォシビルスク)
	Sankt-Peterburg(サンクトペテルブルク)	Ekaterinburg (エカテリンブルク)
Japan (日本)	Tokyo(トーキョー)	Osaka(オオサカ)
	Yokohama(ヨコハマ)	Nagoya(ナゴヤ)
Ethiopia (エチオピア)	Addis Ababa(アジスアベバ)	Gondar(ゴンダール)
	Adaamaa(アダマ)	Maqalle(メックエル)
Egypt (エジプト)	Cairo(カイロ)	Giza(ギザ)
	Alexandria(アレクサンドリア)	Shubra El Kheima(ショブラエルケイマ)
Iran (イラン)	Teheran(テヘラン)	Isfahan(イスファハーン)
	Mashhad(マシュハド)	Karaj(キャラジ)
Dem. Rep. of Congo (コンゴ民主共和国)	Kinshasa(キンシャサ)	Mbuji Mayi(ムブジマイ)
	Lubumbashi(ルブンバシ)	Kisangani(キサンガニ)
British (イギリス)	London(ロンドン)	Glasgow(グラスゴー)
	Birmingham(バーミンガム)	Liverpool(リヴァプール)
Italy (イタリア)	Rome(ローマ)	Naples(ナポリ)
	Milan(ミラノ)	Torino(トリノ)
Myanmar (ミャンマー)	Naypyidaw(ネピドー)	Mandalay(マンダレー)
	Yangon(ヤンゴン)	Mawlamyaing(モーラミャイン)
Korea (韓国)	Incheon(インチョン)	Seoul(ソウル)
	Daegu(テグ)	Busan(プサン)
Spain (スペイン)	Seville(セビリア)	Madrid(マドリード)
	Zaragoza(サラゴサ)	Barcelona(バルセロナ)
Ukraine (ウクライナ)	Odessa(オデッサ)	Kiev(キエフ)
	Dnieper(ドニエプル)	Kharkiv(ハルキウ)
Sudan (スーダン)	Khartoum(ハルツーム)	al Khartum Bahri(アルハルツームバフリ)
	Omdurman(オムドゥルマン)	Nyala(ニャラ)

Uganda (ウガンダ)	Kampala(カンパラ)	Kira(キラ)
	Nansana(ナンサナ)	Makindye(マキンダイ)
Iraq (イラク)	Baghdad(バグダード)	Hillah(ヒッラ)
	Basra(バスラ)	Arbil(アルビル)
Morocco (モロッコ)	Rabat(ラバト)	Fez(フェズ)
	Casablanca(カサブランカ)	Tangier(タンジール)
Saudi Arabia (サウジアラビア)	Riyadh(リヤド)	Mecca(メッカ)
	Jeddah(ジッダ)	Taif(ターイフ)
Venezuela (ベネズエラ)	Caracas(カラカス)	Barquisimeto(バルキシメト)
	Maracaibo(マラカイボ)	Guayana(グアヤナ)
Uzbekistan (ウズベキスタン)	Tashkent(タシケント)	Samarkand(サマルカンド)
	Namangan(ナマンガン)	Andijon(アンディジャン)
Mozambique (モザンビーク)	Maputo(マプト)	Nampula(ナンプラ)
	Matola(マトラ)	Beira(ベイラ)
Yemen (イエメン)	Sanaa(サヌア)	Taiz(タイズ)
	Aden(アデン)	Hudaida(フダイダ)

3. I generated alternative pairs by connecting the correct country (k^{th} highest population) with the country followed by the correct country in the above list (i.e., $[k + 2]^{th}$, $[k + 4]^{th}$, or $[k + 6]^{th}$ highest population). Countries ranked after the 43rd population (Venezuela, Uzbekistan, Mozambique, and Yemen; i.e., $k + 2$, $k + 4$, or $k + 6$ were over 50) were paired with the country ranked 1st, 3rd, 5th, or 7th population (China, America, Brazil, or Nigeria). Some examples of questions are as follows:

"Which country is Beijing in, China (correct; 1st population) or America (3rd population)?"
"Which country is Guangzhou in, China (correct; 1st population) or Brazil (5th population)?"
"Which country is Shanghai in, China (correct; 1st population) or Nigeria (7th population)?"
"Which country is Tianjin in, China (correct; 1st population) or Russia (9th population)?"
"Which country is Washington D.C. in, America (correct; 3rd population) or Brazil (5th population)?"
"Which country is New York in, America (correct; 3rd population) or Nigeria (7th population)?"
"Which country is Los Angeles in, America (correct; 3rd population) or Russia (9th population)?"
"Which country is Chicago in, America (correct; 3rd population) or Japan (11th population)?"

……

"Which country is Nampula in, Mozambique (correct; 47th population) or Yemen (49st population)?"

"Which country is Beira in, Mozambique (correct; 47th population) or China (1st population)?"

"Which country is Maputo in, Mozambique (correct; 47th population) or America (3rd population)?"

"Which country is Matola in, Mozambique (correct; 47th population) or Brazil (5th population)?"

"Which country is Taiz in, Yemen (correct; 49th population) or China (1st population)?"

"Which country is Hudaida in, Yemen (correct; 49th population) or America (3rd population)?"

"Which country is Sana'a in, Yemen (correct; 49th population) or Brazil (5th population)?"

"Which country is Aden in, Yemen (correct; 49th population) or Nigeria (7th population)?"

Note 1: The order in which questions and alternatives were presented was randomized. Of course, participants were not presented the words in parentheses.

Note 2: As described above, I modified the materials used in Study 2 for the following two reasons. First, if I used the same materials as in Study 2, I would force participants to answer too many tasks (50 * 49 binary choice questions). Second, on the other hand, if the number of tasks was too small, I would not be able to apply model selection analyses because analyses based on cognitive models generally need many data. To avoid participants' overwork but to obtain enough data for conducting model analyses, I picked parts of the materials of Study 2 in the above way.

S.5 Supplementary material 5

The hit numbers (bold font in the right column) of odd-ranked and even-ranked 125 objects used in Study 4.

Note: "Rank" denotes the country's population size (searched in May 2016). From each country (i.e., odd-ranked country 1st, 3rd, …, 49th; even-ranked country 2nd, 4th, …, 50th, in the population size), four cities were picked up according to the highest population. Data accessed for hit numbers of odd-ranked objects at 06/11/2020, and for hit numbers of even-ranked objects at 07/05/2020, in an online-newspaper database, *Maisaku* (provided by a Japanese newspaper, Mainichi-shimbun).

Odd rank data (100 cities and 25 countries)

category	Rank	English description	Japanese description	**Hit number**
city	1	Beijing	ペキン	**17671**
city	1	Shanghai	シャンハイ	**6393**
city	1	Guangzhou	コウシュウ	**1820**
city	1	Tianjin	テンシン	**757**
city	3	Washington D.C.	ワシントン D.C.	**18205**
city	3	Los Angeles	ロサンゼルス	**3049**
city	3	New York	ニューヨーク	**11780**
city	3	Chicago	シカゴ	**1117**
city	5	Brasilia	ブラジリア	**230**
city	5	Rio de Janeiro	リオデジャネイロ	**1677**
city	5	Sao Paulo	サンパウロ	**1109**
city	5	Salvador	サルヴァドール	**3**
city	7	Abuja	アブジャ	**53**
city	7	Ibadan	イバダン	**4**
city	7	Lagos	ラゴス	**43**
city	7	Kano	カノ	**25**
city	9	Moscow	モスクワ	**7045**
city	9	Novosibirsk	ノヴォシビルスク	**22**

city	9	Sankt-Peterburg	サンクトペテルブルク	**787**
city	9	Ekaterinburg	エカテリンブルク	**155**
city	11	Tokyo	トーキョー	**152084**
city	11	Osaka	オオサカ	**62128**
city	11	Yokohama	ヨコハマ	**17087**
city	11	Nagoya	ナゴヤ	**19061**
city	13	Addis Ababa	アジスアベバ	**116**
city	13	Gondar	ゴンダール	**3**
city	13	Adaamaa	アダマ	**0**
city	13	Maqalle	メックエル	**0**
city	15	Cairo	カイロ	**2110**
city	15	Giza	ギザ	**105**
city	15	Alexandria	アレクサンドリア	**177**
city	15	Shubra El Kheima	ショブラエルケイマ	**1**
city	17	Teheran	テヘラン	**1894**
city	17	Isfahan	イスファハーン	**69**
city	17	Mashhad	マシュハド	**6**
city	17	Karaj	キャラジ	**5**
city	19	Kinshasa	キンシャサ	**57**
city	19	Mbuji Mayi	ムブジマイ	**1**
city	19	Lubumbashi	ルブンバシ	**1**
city	19	Kisangani	キサンガニ	**0**
city	21	London	ロンドン	**6391**
city	21	Glasgow	グラスゴー	**373**
city	21	Birmingham	バーミンガム	**225**
city	21	Liverpool	リヴァプール	**200**
city	23	Rome	ローマ	**2928**
city	23	Naples	ナポリ	**796**
city	23	Milan	ミラノ	**1916**
city	23	Torino	トリノ	**884**
city	25	Naypyidaw	ネピドー	**331**
city	25	Mandalay	マンダレー	**45**
city	25	Yangon	ヤンゴン	**1027**
city	25	Mawlamyaing	モーラミャイン	**3**
city	27	Incheon	インチョン	**1385**

city	27	Seoul	ソウル	**10562**
city	27	Daegu	テグ	**686**
city	27	Busan	プサン	**1712**
city	29	Seville	セビリア	**688**
city	29	Madrid	マドリード	**2456**
city	29	Zaragoza	サラゴサ	**365**
city	29	Barcelona	バルセロナ	**2356**
city	31	Odessa	オデッサ	**35**
city	31	Kiev	キエフ	**654**
city	31	Dnieper	ドニエプル	**17**
city	31	Kharkiv	ハルキウ	**71**
city	33	Khartoum	ハルツーム	**62**
city	33	al Khartum Bahri	アルハルツームバフリ	**0**
city	33	Omdurman	オムドゥルマン	**1**
city	33	Nyala	ニャラ	**0**
city	35	Kampala	カンパラ	**38**
city	35	Kira	キラ	**0**
city	35	Nansana	ナンサナ	**1**
city	35	Makindye	マキンダイ	**0**
city	37	Baghdad	バグダード	**1309**
city	37	Hillah	ヒッラ	**19**
city	37	Basra	バスラ	**135**
city	37	Arbil	アルビル	**200**
city	39	Rabat	ラバト	**25**
city	39	Fez	フェズ	**5**
city	39	Casablanca	カサブランカ	**37**
city	39	Tangier	タンジール	**4**
city	41	Riyadh	リヤド	**276**
city	41	Mecca	メッカ	**135**
city	41	Jeddah	ジッダ	**115**
city	41	Taif	ターイフ	**2**
city	43	Caracas	カラカス	**176**
city	43	Barquisimeto	バルキシメト	**1**
city	43	Maracaibo	マラカイボ	**1**
city	43	Guayana	グアヤナ	**0**

city	45	Tashkent	タシケント	**170**
city	45	Samarkand	サマルカンド	**21**
city	45	Namangan	ナマンガン	**2**
city	45	Andijon	アンディジャン	**8**
city	47	Maputo	マプト	**19**
city	47	Nampula	ナンプラ	**2**
city	47	Matola	マトラ	**3**
city	47	Beira	ベイラ	**3**
city	49	Sanaa	サヌア	**262**
city	49	Taiz	タイズ	**35**
city	49	Aden	アデン	**121**
city	49	Hudaida	フダイダ	**1**
country	1	China	中国	**101450**
country	3	America	アメリカ	**123460**
country	5	Brazil	ブラジル	**13368**
country	7	Nigeria	ナイジェリア	**1719**
country	9	Russia	ロシア	**37529**
country	11	Japan	日本	**475927**
country	13	Ethiopia	エチオピア	**1686**
country	15	Egypt	エジプト	**9082**
country	17	Iran	イラン	**12700**
country	19	DemRepCongo	コンゴ民主共和国	**537**
country	21	British	イギリス	**28424**
country	23	Italy	イタリア	**19183**
country	25	Myanmar	ミャンマー	**4464**
country	27	Korea	韓国	**55034**
country	29	Spain	スペイン	**15641**
country	31	Ukraine	ウクライナ	**6779**
country	33	Sudan	スーダン	**943**
country	35	Uganda	ウガンダ	**594**
country	37	Iraq	イラク	**11198**
country	39	Morocco	モロッコ	**1316**
country	41	Saudi Arabia	サウジアラビア	**4295**
country	43	Venezuela	ベネズエラ	**2215**
country	45	Uzbekistan	ウズベキスタン	**1578**

country	47	Mozambique	モザンビーク	**240**
country	49	Yemen	イエメン	**1447**

Even rank data (100 cities and 25 countries)

Category	Rank	English description	Japanese description	**Hit number**
City	2	Delhi	デリー	**1657**
City	2	Mumbai	ムンバイ	**492**
City	2	Kolkata	コルカタ	**107**
City	2	Chennai	チェンナイ	**92**
City	4	Jakarta	ジャカルタ	**1504**
City	4	Surabaya	スラバヤ	**76**
city	4	Bandung	バンドン	**100**
city	4	Medan	メダン	**29**
city	6	Karachi	カラチ	**206**
city	6	Lahore	ラホール	**121**
city	6	Faisalabad	ファイサラバード	**4**
city	6	Quetta	クエッタ	**62**
city	8	Dhaka	ダッカ	**349**
city	8	Chattogram	チッタゴン	**32**
city	8	Khulna	クルナ	**2**
city	8	Narayanganj	ナラヤンガンジ	**1**
city	10	Mexico City	メキシコシティ	**1165**
city	10	Ecatepec	エカテペック	**0**
city	10	Guadalajara	グアダラハラ	**89**
city	10	Puebla	プエブラ	**18**
city	12	Quezon	ケソンシティ	**19**
city	12	Manila	マニラ	**1266**
city	12	Caloocan	カローカン	**0**
city	12	Davao	ダバオ	**121**
city	14	Ho Chi Minh	ホーチミン	**509**

city	14	Hanoi	ハノイ	**875**
city	14	Haiphong	ハイフォン	**25**
city	14	Da Nang	ダナン	**134**
city	16	Berlin	ベルリン	**3079**
city	16	Hamburg	ハンブルク	**365**
city	16	Munich	ミュンヘン	**1703**
city	16	Cologne	ケルン	**675**
city	18	Istanbul	イスタンブル	**1059**
city	18	Ankara	アンカラ	**342**
city	18	Izmir	イズミル	**43**
city	18	Bursa	ブルサ	**7**
city	20	Bangkok	バンコク	**873**
city	20	Nonthaburi	ノンタブリー	**3**
city	20	Nakhon Ratchasima	ナコーンラーチャシーマー	**1**
city	20	Chiang Mai	チェンマイ	**54**
city	22	Paris	パリ	**8474**
city	22	Marseilles	マルセイユ	**840**
city	22	Lyon	リヨン	**928**
city	22	Toulouse	トゥールーズ	**598**
city	24	Cape Town	ケープタウン	**236**
city	24	Durban	ダーバン	**289**
city	24	Johannesburg	ヨハネスブルク	**758**
city	24	Soweto	ソウェト	**57**
city	26	Dar es Salaam	ダルエスサラーム	**26**
city	26	Mwanza	ムワンザ	**1**
city	26	Arusha	アルーシャ	**2**
city	26	Dodoma	ドドマ	**0**
city	28	Bogotá	ボゴタ	**79**
city	28	Medellín	メデジン	**58**
city	28	Cali	カリ	**36**
city	28	Barranquilla	バランキージャ	**2**
city	30	Nairobi	ナイロビ	**421**
city	30	Mombasa	モンバサ	**57**
city	30	Kisumu	キスム	**25**
city	30	Nakuru	ナクル	**17**

city	32	Buenos Aires	ブエノスアイレス	**424**
city	32	Rosario	ロサリオ	**23**
city	32	Mendoza	メンドーサ	**1**
city	32	San Miguel de Tucumán	トゥクマン	**3**
city	34	Algiers	アルジェ	**96**
city	34	Oran	オラン	**11**
city	34	Constantine	コンスタンティーヌ	**0**
city	34	al-Gilfah	ジェルファ	**0**
city	36	Warsaw	ワルシャワ	**303**
city	36	Kraków	クラクフ	**74**
city	36	Lodz	ウッチ	**17**
city	36	Wrocław	ブレスラウ	**11**
city	38	Toronto	トロント	**685**
city	38	Montreal	モントリオール	**494**
city	38	Calgary	カルガリー	**410**
city	38	Ottawa	オタワ	**210**
city	40	Kabul	カブール	**1031**
city	40	Kandahar	カンダハール	**186**
city	40	Herat	ヘラート	**58**
city	40	Mazare Sharīf	マザリシャリフ	**27**
city	42	Lima	リマ	**249**
city	42	Arequipa	アレキパ	**12**
city	42	Trujillo	トルヒーリョ	**1**
city	42	Chiclayo	チクラヨ	**1**
city	44	Kuala Lumpur	クアラルンプール	**860**
city	44	Johor Bahru	ジョホールバル	**43**
city	44	Ipoh	イポー	**9**
city	44	Subang Jaya	スバンジャヤ	**1**
city	46	Kathmandu	カトマンズ	**766**
city	46	Pokhara	ポカラ	**65**
city	46	Lalitpur	ラリトプル	**18**
city	46	Biratnagar	ビラートナガル	**5**
city	48	Accra	アクラ	**64**
city	48	Kumasi	クマシ	**3**
city	48	Tamale	タマレ	**4**

city	48	Sekondi-Takoradi	セコンディタコラディ	**0**
city	50	Pyeongyang	ピョンヤン	**2705**
city	50	Hamhung	ハムン	**73**
city	50	Chongjin	チョンジン	**124**
city	50	Sinuiju	シニジュ	**80**
country	2	India	インド	**24151**
country	4	Indonesia	インドネシア	**7065**
country	6	Pakistan	パキスタン	**4635**
country	8	Bangladesh	バングラデシュ	**1504**
country	10	Mexicanos	メキシコ	**9335**
country	12	Philippines	フィリピン	**7671**
country	14	Vietnam	ベトナム	**8138**
country	16	Germany	ドイツ	**32144**
country	18	Turkey	トルコ	**7305**
country	20	Thailand	タイ	**9070**
country	22	France	フランス	**30705**
country	24	South Africa	南アフリカ	**7708**
country	26	Tanzania	タンザニア	**500**
country	28	Colombia	コロンビア	**2542**
country	30	Kenya	ケニア	**3262**
country	32	Argentine	アルゼンチン	**4813**
country	34	Algeria	アルジェリア	**1559**
country	36	Poland	ポーランド	**4775**
country	38	Canada	カナダ	**18698**
country	40	Afghanistan	アフガニスタン	**7990**
country	42	Peru	ペルー	**3565**
country	44	Malaysia	マレーシア	**6268**
country	46	Nepal	ネパール	**3984**
country	48	Ghana	ガーナ	**1480**
country	50	North Korea	北朝鮮	**32725**

S.6 Supplementary material 6

The objects which were used in Study 5's binary choice task (consumer choices). The question was "which item do you want to buy?" and the alternative format was "Q made by company A (or B)." Words in parentheses are Japanese descriptions, which were presented to participants in Study 5.

Category	**Situation**	**Alt. A (familiar company)**	**Alt. B (unfamiliar company)**
Alcohol	Q1 (familiar item)	Beer made by Suntory(サントリー社製のビール)	Beer made by Torei(トーレイ社製のビール)
	Q2 (unfamiliar item)	Tokaji made by Suntory(サントリー社製のトカイ)	Tokaji made by Torei(トーレイ社製のトカイ)
Season-ing	Q3 (familiar item)	Soy sauce made by Kikkoman(キッコーマン社製の醤油)	Soy sauce made by Hung Thanh(フンタン社製の醤油)
	Q4 (unfamiliar item)	Nuoc mam made by Kikkoman(キッコーマン社製のニョクマム)	Nuoc mam made by Hung Thanh(フンタン社製のニョクマム)
Tea	Q5 (familiar item)	Green tea made by Ito En(伊藤園社製の緑茶)	Green tea made by Premier's Tea(プリミアスティー社製の緑茶)
	Q6 (unfamiliar item)	Nilgiri made by Ito En(伊藤園社製のニルギリ)	Nilgiri tea made by Premier's Tea(プリミアスティー社製のニルギリ)
Cup	Q7 (familiar item)	Cup made by Kyocera(京セラ社製のコップ)	Cup made by Moser(モーゼル社製のコップ)
	Q8 (unfamiliar item)	Bohemian glass made by Kyocera(京セラ社製のボヘミアングラス)	Bohemian glass made by Moser(モーゼル社製のボヘミアングラス)
Bicycle	Q9 (familiar item)	Bicycle made by Bridgestone(ブリヂストン社製の自転車)	Bicycle made by Cannondale(キャノンデール社製の自転車)
	Q10 (unfamiliar item)	Road bike made by Bridgestone(ブリヂストン社製のロードバイク)	Road bike made by Cannondale(キャノンデール社製のロードバイク)
Pen	Q11 (familiar item)	Ball point pen made by Tombo(トンボ社製のボールペン)	Ball point pen made by Lamy(ラミー社製のボールペン)

	Q12 (unfamiliar item)	Fountain pen made by Tombo(トンボ社製の万年筆)	Fountain pen made by Lamy(ラミー社製の万年筆)
Camera	Q13 (familiar item)	Compact camera made by Fujifilm(富士フィルム社製のコンパクトカメラ)	Compact camera made by Hasselblad(ハッセルブラッド社製のコンパクトカメラ)
	Q14 (unfamiliar item)	Digital single-lens reflex camera made by Fujifilm(富士フィルム社製のデジタル一眼レフレックスカメラ)	Digital single-lens reflex camera made by Hasselblad(ハッセルブラッド社製のデジタル一眼レフレックスカメラ)

Note: The order of questions (Q1 ~ Q14) and alternatives (alt. A, and alt. B) was randomized. I assumed that, in one question, one company would be familiar and the other unfamiliar to many Japanese people. I also assumed that, in one category, one item would be familiar and the other unfamiliar to Japanese people.

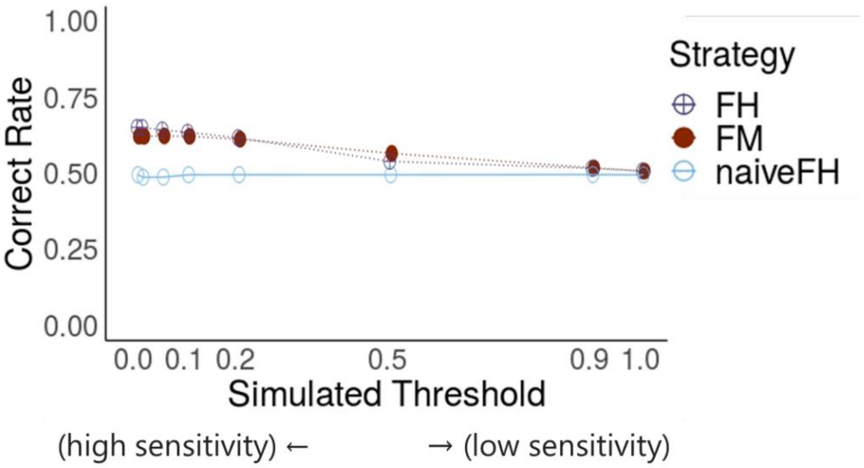
S.7 Supplementary material 7

The comparison of performances between familiarity-matching (FM), familiarity heuristic (FH), and another simpler model (naïve FH) through computer simulations

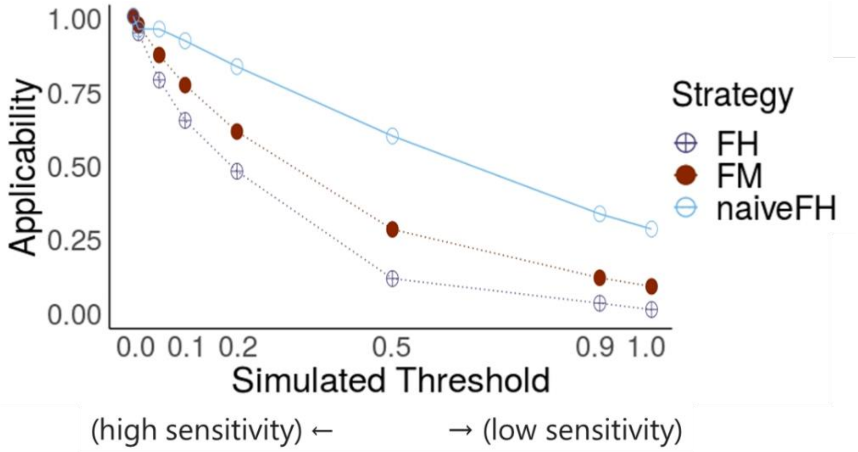
Based on the discussion about the results on the applicability of FM in Study 4b, one may argue that FM is simply more favorable to a relationships-comparison task than FH, in terms of the number of familiarity which is compared with the decision threshold. In FM, a decision threshold is compared with only one term: "||FamQ – FamA| – |FamQ – FamB||." In FH, on the other hand, it is compared with two terms: "|FamQ – medianFamQs|" and "|FamA – FamB|" (see also Chapter 2) Because the computational condition of FH is stricter than that of FM, it is natural that using FH requires a lower decision threshold than using FM. To address this issue, I introduced another familiarity heuristic model which was equivalent to FM in terms of the number of to-be-compared familiarity. I assumed that people constantly choose the recognized or more familiar alternative regardless of FamQ, as proposed in previous studies (e.g., Goldstein & Gigerenzer, 2002; Honda et al., 2011, 2017), and call this model "naïve FH." In naïve FH, as in FM, a decision threshold is compared with only one term: "|FamA– FamB|."

As to naïve FH, I conducted computer simulations with the same procedure as in Study 4b, and then compared performances of naïve FH with those of FM and FH. I found that, although naïve FH had the higher applicability than FM and FH (i.e., people had more chances to use naïve FH in the task), it showed the lower correct rate than FM and FH. These findings indicate that naïve FH can be applied to many cases but is less likely to lead correct inferences. It may be because naïve FH is assumed to consider only two countries A and B but not to consider a city Q. That is, naïve FH does not compare two relationships (i.e., "city Q and country A" and "city Q and country B") in a relationships-comparison task. Based on these results, FM is considered to be an advantageous and adaptive heuristic for a relationships-comparison task in terms of both the number of familiarity which is compared with people's decision threshold, and the comparison of relationships between objects (these two contribute to the high applicability and the high correct rate, respectively).

The correct rate and applicability of naïve familiarity heuristic (naïve FH) (light-blue line), which assumes that people constantly choose the recognized or more familiar alternative. For comparisons, the applicability and correct rate of familiarity-matching (FM) and of familiarity heuristic (FH) were also shown by dotted lines (the same data as in Fig. 16). The x-axis denotes agents' decision threshold for discriminating between similarities in familiarity, as in Fig. 16 (i.e., the left [right] on the x-axis denotes higher [lower] sensitivity).



(A)



(B)

S.8 Supplementary material 8

The distributions and effects of "relationships"

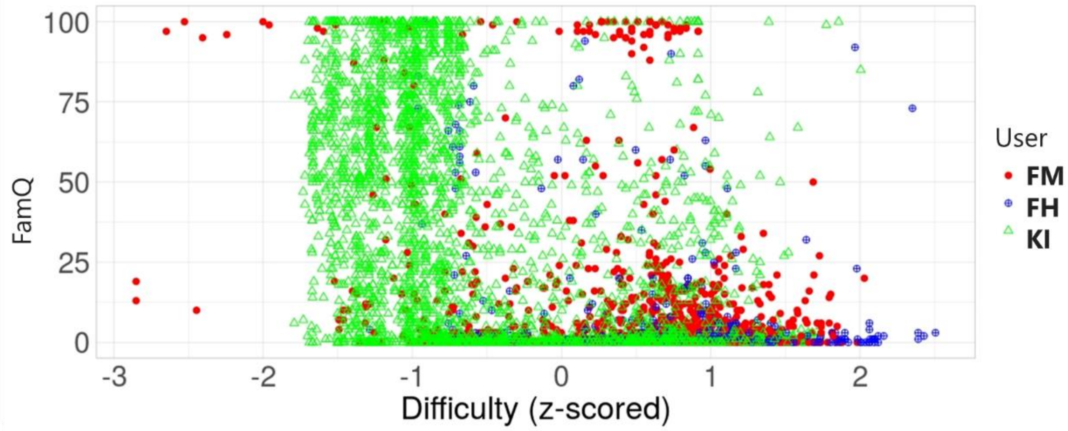
Note: In this section, the following abbreviations are used.
FM: familiarity-matching; *FH*; familiarity heuristic; *KI (or Knowledge)*: knowledge-based inference; *Q (or object Q)*: the object that is presented in a question; *FamQ*: familiarity of object Q; *FamC*: familiarity of the correct alternative; *FamF*: familiarity of the false alternative; *KnowQ*: the number of recognized attributes about object Q; *KnowMatchQC*: the number of recognized attributes that accord between object Q and the correct alternative; *KnowMatchQF*: the number of recognized attributes that accord between object Q and the false alternative.

To investigate features of environments that will make us use FM, further exploratory analyses on the difficulty and the use of FM were conducted, using Study 3's data. In what environments were people (i.e., FM users, FH users, and KI users) likely to feel more difficulty? And what environments would enhance the use of FM? In a relationships-comparison task, since it would be important to consider relationships between the subjective difficulty, a city in a question and the correct/false countries (e.g., the correlation between difficulty and familiarity, |FamQ − FamC|, and KnowMatchQC, etc.), I focused on familiarity and knowledge of object Q, and relationships between object Q and alternatives (i.e., differences between familiarity of Q and that of the alternative, and the number of accorded attributes between Q and the alternative, etc.).

In the following analyses, the difficulty ratings provided from participants were transformed into z-scores for each participant. Although no remarkable tendencies were observed, I will report these results.

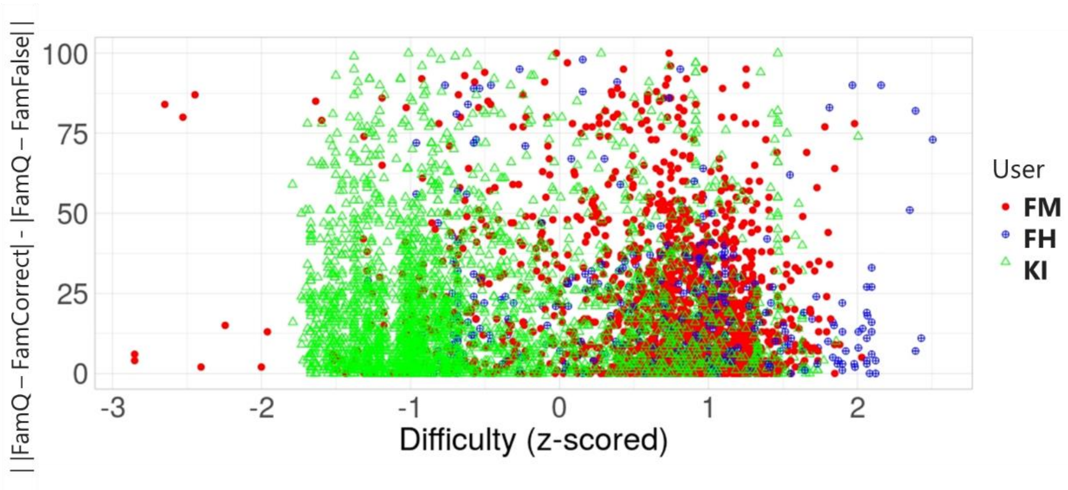
Relationship between the difficulty and FamQ

(i.e., an examination of how much the difficulty would be related to familiarity of object Q)

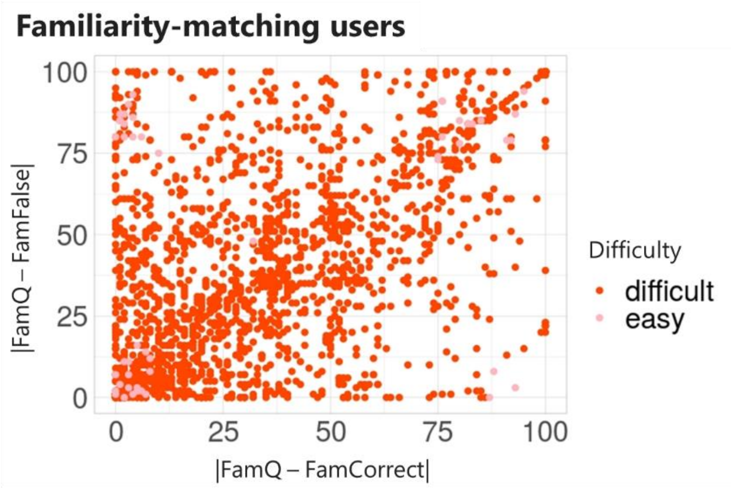


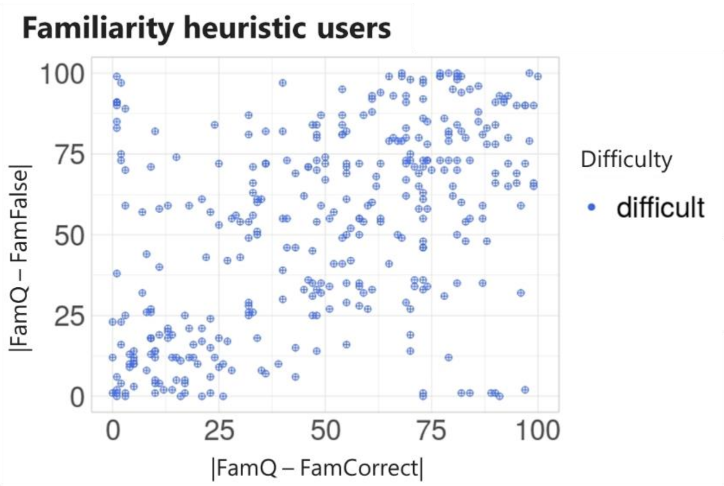
Relationship between the difficulty and ||FamQ – FamC – |FamQ – FamF||

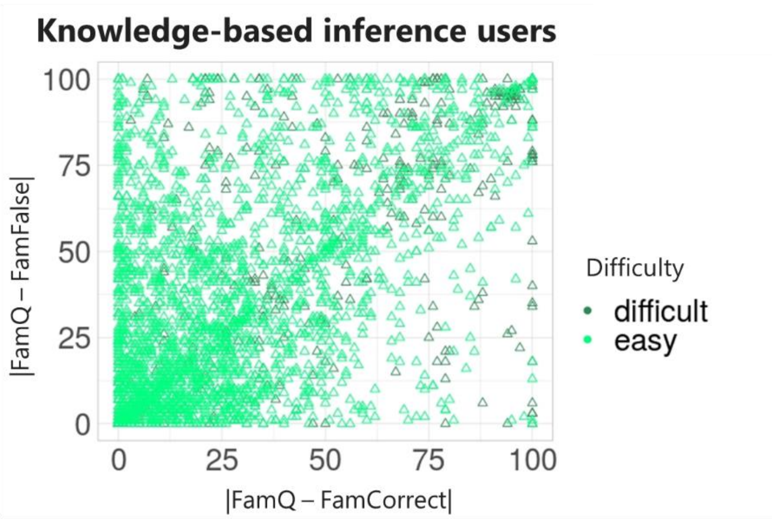
(i.e., an examination of how much the difficulty would be related to familiarity that is considered in using FM)



Distributions of familiarity for each user (FM user, FH user, and KI user)







Next, I predicted the difficulty and the use of FM using linear mixed models. In the following analyses, a random effect of linear mixed models was each participant's decision threshold, and familiarity ratings were transformed into z-scored for each participant. In order to focus on cases where heuristics were likely to be used (i.e., participants did not know the correct answer), I excluded cases where participants made correct inferences for "country" questions in the knowledge task (i.e., "Which country do you think that the city is in?"). Then, I analyzed the remained 2,589 cases. As described below, although the main focuses were the "relationships" between Q and alternatives (i.e., |FamQ – FamC|, |FamQ – FamF|, KnowMatchQC, and KnowMatchQF), the effects of only Q (i.e., FamQ and KnowQ) were introduced as fixed effects of linear mixed models. In a relationships-comparison task, since participants first would look at Q in a question and then they would see alternatives and consider these relationships, I assumed that FamQ and KnowQ might strongly affect participants' inferences. Thus, not only |FamQ – FamC|, |FamQ – FamF|, KnowMatchQC, and KnowMatchQF, but also FamQ and KnowQ were introduced as fixed effects.

Prediction of difficulty (by a linear mixed model)

Dependent variable: difficulty ratings provided from participants

Fixed effects: KnowQ, FamQ, KnowMatchQC, KnowMatchQF, |FamQ – FamC|, and |FamQ – FamF|

	Estimate	Std. Error	Df	*t*-value	*p*-value
(intercept)	0.45788	0.06113	116.13836	7.490	< .001
KnowQ	0.03228	0.01560	2518.62644	2.070	< .05
FamQ	-0.39971	0.03243	2581.9257	-12.325	< .001
KnowMatchQC	-0.05557	0.02068	2553.60336	-2.687	< .01
KnowMatchQF	0.04791	0.02173	2551.9964	2.205	< .05
|FamQ – FamC|	-0.07248	0.02359	2565.67636	-3.073	< .01
|FamQ – FamF|	-0.23090	0.01912	2563.7203	-12.077	< .001

Prediction of the use of FM (by a generalized linear mixed model)

Dependent variable: a dummy (whether FM was used [1] or not [0])

Fixed effects: FamQ, KnowQ, KnowMatchQC, KnowMatchQF, and interaction terms between these effects and difficulty

	Estimate	Std. Error	*z*-value	*p*-value
(intercept)	-6.10484	2.25120	-2.712	< .01
KnowQ	0.42209	0.12384	3.408	< .001
FamQ	-0.67357	0.21118	-3.189	< .01
KnowMatchQC	-0.64793	0.17383	-3.727	< .001
KnowMatchQF	0.08703	0.21562	0.404	.69
|FamQ – FamC|	-0.77183	0.14612	-5.282	< .001
|FamQ – FamF|	-0.84155	0.12664	-6.645	< .001
KnowQ * difficulty	-0.07475	0.12377	-0.604	.55
FamQ * difficulty	0.69106	0.21004	3.290	< .01
KnowMatchQC * difficulty	0.48851	0.18052	2.706	< .01
KnowMatchQF * difficulty	-0.30276	0.21796	-1.389	.16
|FamQ – FamC| * difficulty	0.21191	0.15440	1.372	.17
|FamQ – FamF| * difficulty	0.24160	0.12929	1.869	.06

S.9 Supplementary material 9

The additional analyses on "sparse" raters (Study 1's data)

There might be participants who rated familiarity sparsely, not continuously (e.g., only 2-points rating such as "0 or 100"), even though they were asked to rate it in a visual analog scale. If so, the assumptions of heuristic models might be inappropriate. In order to consider these issues, I reported the results of additional analyses on such "sparse" raters.

For each difficulty level, I calculated the mean of the number of rating values (e.g., if a participant rated only 0 or 100, then the number of rating values is "2") in the difficulty rating task and in the measurement of familiarity task for each participant. That is, it means that the lower this value was, the more sparsely the participant rated difficulty and familiarity. In addition, if a participant's the mean of the number of rating value was below 2SD, then I defined the participant as "sparse" raters and the others as "non-sparse" raters. Then, I show the means of the number of rating values in classified and not-classified participants, and the values of G^2 (an index for the goodness of fit of a certain model; see section 3.2.1) for FM and FH in sparse and non-sparse raters. Note that, I could not conduct statistical tests because of extremely skewed sample sizes in the current analyses.

The means of the number of rating values
Note: The words “Classified” and “Not classified” mean that participants who were and were not classified into a certain model in model selection analyses, respectively. For details, see Study 1.

The results imply that both classified participants and not-classified participants did not tend to rate difficulty and familiarity sparsely. So, it is unlikely that not-classified participants were not classified into a certain model due to their sparse ratings.

		Median	Mean	SD
Difficult	All ($n = 90$)	18.6	17.5	5.31
	Classified ($n = 81$)	19.0	17.8	4.99
	Not classified ($n = 9$)	18.0	14.5	7.34
Easy	All ($n = 90$)	21.5	21.2	5.28
	Classified ($n = 89$)	21.8	21.3	5.31
	Not classified ($n = 1$)	20.8	20.8	---

The values of G^2 of “sparse” raters and “non-sparse” raters
Note: G^2 is an index for the goodness of fit of a certain model. For details, see Study 1.

The results imply that the values did not so much differ between sparse and non-sparse raters. So, it is unlikely that the model-fitting performances drastically decrease due to sparse ratings.

			Median	Mean	SD
Difficult	FM	“Sparse” ($n = 1$)	83.2	83.2	---
		“Non-sparse” ($n = 80$)	78.2	76.9	4.56
	FH	“Sparse” ($n = 1$)	83.2	83.2	---
		“Non-sparse” ($n = 80$)	77.7	77.6	3.82
Easy	FM	“Sparse” ($n = 5$)	81.8	80.8	2.72
		“Non-sparse” ($n = 84$)	80.4	79.8	2.44
	FH	“Sparse” ($n = 5$)	81.8	79.2	4.97
		“Non-sparse” ($n = 84$)	77.4	76.0	5.35

Footnotes

[i] To the best of my knowledge, there is no clear definition of "environment" or "environmental structure" in this field, as described in the beginning of this thesis. In fact, Gigerenzer et al. (1999) described as follows: "We do not yet have a well-developed language for describing those aspects of environment structure, whether physical or social, that shape the design and performance of decision heuristics. (p. 364)" In this thesis, I would like to agree with and adopt Clark (2010)'s description (Clark, 2010) "the environment is the world around us, where we act, what we are affected by and, in turn, that which is affected by our actions", and Gigerenzer's description (in Chater et al., 2018) "The term environment, as defined by the ecological axioms, does not relate to a world independent of humans, …but to the world as experienced by humans". Based on these descriptions, I provide the definitions of these terms in this thesis; see a preface of Chapter 1.

[ii] Typically, human choice behaviors are categorized into *preferences* (i.e., situations without an objective criterion for deciding which alternative has a higher or lower value; such as a matter of tastes) or *inferences* (i.e., situations with such an objective criterion), and these two draw on the same cognitive processes (e.g., Weber & Johnson, 2009). However, ecological rationality of heuristics has been evaluated by using inferential tasks. This is because inferential tasks have a unique criterion (in other words, an objectively verifiable answer) and thus researchers can quantify the rationality by calculating the correct rate (i.e., accuracy) (e.g., Gigerenzer & Gaissmaier, 2011).

[iii] Although the fluency heuristic (e.g., Hertwig, et al., 2008; Schooler & Hertwig, 2005) is one of the well-known heuristics, I did not examine it in this thesis for the following two reasons. First, the fluency heuristic is highly difficult to disentangle from the familiarity heuristic in a theoretical aspect (Honda et al., 2017). Second, Honda et al. (2017) demonstrated that the familiarity heuristic model could explain participants' inferential patterns better than the fluency heuristic model, through model selection analyses. Because the familiarity heuristic model and knowledge-based inference model could most effectively explain human inferential patterns in difficult and easy questions, respectively (Honda et al., 2017), hereafter in this thesis I focus on the familiarity and knowledge as people's inferential cues.

[iv] Like the term "environment," there are no clear definitions of "cognition" and "context" aspects in Simon's scissors metaphor. I refer to Clark (2010)'s two descriptions (as in the original sources): "Cognition is "the act of thinking"…. This statement is intended to be descriptive rather than prescriptive.", and "Context is typically used in two related ways. In many respects, it is simply short-hand for the environment. In some cases, though, context has a narrower meaning referring to the specific domain in which behaviour takes place". Based on them, I will refer a "cognition" blade as the way of human inferences, and a "context" blade as a task structure which will relate to an environmental structure.

[v] As described in Chapter 1, I used this format in Studies 1~4 as an inferential task which focused on the frequency of appearance of information in the real world (i.e., general knowledge that people will see or hear through media). In Study 5, on the other hand, I used another format as a preference task (i.e., people are not asked about general knowledge; but the task structure was identical to the task in Study 1~4). For details of the task in Study 5, see Study 5.

[vi] Although both FM and FH are inference models based on the familiarity of objects and may seem almost identical to each other, the prediction of FM will sometimes differ from that of FH. For example, consider the following situation in which two questions are presented. In question 1, $FamQ_1 = 90$, $FamA_1 = 60$, and $FamB_1 = 10$, and, in question 2, $FamQ_2 = 40$, $FamA_2 = 60$, and $FamB_2 = 10$. In this situation, FM predicts that people will choose alternative A in both questions because both $FamQ_1$ and $FamQ_2$ are more similar to familiarity of alternative A ($|FamQ_1 - FamA_1| = 30 < |FamQ_1 - FamB_1| = 80$ in question 1; $|FamQ_2 - FamA_2| = 20 < |FamQ_2 - FamB_2| = 30$ in question 2). On the other hand, FH predicts that people choose alternative A in question 1 but choose alternative B in question 2; because $FamQ_1 > medianFamQs$ (= 65), they choose the more familiar alternative while, because $FamQ_2 < medianFamQs$, they choose the more unfamiliar one. Even if FamA and FamB are the same between questions, FM and FH's predictions can differ from each other depending on FamQ.

[vii] One may argue that using phrases, such as "know much," seemed to refer to knowledge and not familiarity. However, as previous studies (e.g., Brown & Siegler, 1992, 1993; Honda et al., 2011, 2017) suggested, I believe that knowledge and familiarity differ from each other as far as knowledge represents the amount of content on an object and familiarity represents the degree of exposure to the object. Thus, I regard the descriptions for the ratings of familiarity (e.g., "know much"; in fact, Honda et al. used the same descriptions as in this thesis) as familiarity, and participants' responses for questions about a certain attribution (e.g., "religion") as knowledge.

[viii] As examples for calculations of G^2, I show the following three cases: more correct application case; less correct application; and all guessing cases (hereafter, assuming a total eight questions for simplification). More correct application case: A participant guessed on two questions, applied a certain inference strategy to five questions, and did not apply the strategy to one question. In such a case, $\varepsilon = 1 / (8-2) = .17$ and $G^2 = -2 * \{\ln(.50) + \ln(.50) + \ln(.83) + \ln(.83) + \ln(.83) + \ln(.17)\} = 7.81$. Less correct application case: A participant guessed on two questions, applied a certain strategy to one question, and did not apply it to five questions. In this case, $\varepsilon = 5 / (8-2) = .17$ and $G^2 = -2 * \{\ln(.50) + \ln(.50) + \ln(.83) + \ln(.17) + \ln(.17) + \ln(.17)\} = 17.47$. All guessing case: If a participant guessed on all eight questions (an "all guessing" case), then $f_x(y) = .50$ and $G^2 = -2 * \{\ln(.50) + \ln(.50) + \ln(.50) + \ln(.50) + \ln(.50) + \ln(.50)\} = 9.70$. Thus, the lower a value of G^2 was, the more correctly s/he applied the strategy to inference tasks (i.e., the better model fit).

[ix] I confirmed the robustness of the FH model by using the 1st or 3rd quantile of FamQs rather than the median of FamQs. In both difficult and easy questions, I found that the accordance rates of the prediction between FH with the median and FH with the 1st or 3rd quantile were very high (difficult: "FH with the median -- FH with the 1st quantile": .99; "FH with the median -- FH with the 3rd quantile": .92. easy: "FH with the median -- FH with the 1st quantile": .94; "FH with the median -- FH with the 3rd quantile": .99). These results indicate that the obtained results on the FH model can be considered robust even when I use the quantiles of FamQs instead of the medianFamQs.

[x] To confirm this issue, using the data of Study 1, I assumed that participants constantly chose the more familiar alternative (i.e., familiarity heuristic that has originally been proposed by Honda et al. (2017); identical to "naïve FH" model in Study 4b's discussion in this thesis) and then calculated the mean correct rates for both difficult and easy questions. These correct rates were .41 and .53, respectively, and were not significantly different from the chance level, .50 (difficult: $W = 8$, $p = .16$, $d = 0.15$; easy $W = 65.5$, $p = .43$, $r = 0.08$, Wilcoxon rank sum test). Thus, it suggests that a heuristic based only on objects comparison such as naïve FH can become less adaptive in a relationships-comparison task.

[xi] It was because the standard deviations of the numbers of hits were very large ($M_{cities} = 3.5 * 10^4$, $SD_{cities} = 1.8 * 10^5$; $M_{countries} = 7.0 * 10^4$, $SD_{countries} = 2.1 * 10^5$) and there were some outliers (e.g., in 50 cities, the highest hit number [for "Tokyo"] was $1.2 * 10^6$, and the second highest number [for "Washington D. C."] was $7.1 * 10^4$).

[xii] Generally, to conduct model-based analyses, cognitive modelers need as many experimental data as possible. Also in this study, actually I should conduct many binary choice tasks and should have to assume both situations in which participants would use heuristics (difficult questions, i.e., countries with low population) and situations in which they would use knowledge (easy questions, i.e., countries with high population), as in Study 1. However, if I used all objects and generated all possible pairs for the behavioral experiment, participants would have had to answer too many tasks (e.g., 225 objects [50 countries and 4 cities * 50 countries = 200 cities] and 50 * 49 * 0.5 = 1225 alternative pairs). Therefore, I picked up only 125 objects (25 countries and 100 cities) for this procedure.

[xiii] Because there were much more FM users than FH users, one could argue that the binary choice task in Study 3 unexpectedly consisted of a question set in which FM users could choose correct answers more easily than FH users. However, correct rates for the binary choice task did not so differ between them in difficult questions (FM users .61 and FH users .66, not significant; $W = 62.5$, $p = .12$, $r = 0.24$, Wilcoxon rank sum test). Therefore, the questions would not necessarily be more advantageous to FM users than to FH users.

[xiv] https://mainichi.jp/contents/edu/maisaku/. I input a country name only in searching a country (e.g., "India"). However, I input a city name and the corresponding country name in searching a city (e.g., "India Delhi"). It was because, when some city names were looked up, an identical city name in a different country (e.g., the city "Cali" ["カリ" in Japanese], which is the third largest city in Columbia, also exists in Switzerland) or many irrelevant results (e.g., in searching "Paris" ["パリ" in Japanese], many irrelevant words such as "sappari" ["サッパリ" in Japanese, which means "refreshing"] appeared in search results) sometimes came up. Note that all words were input in Japanese (e.g., "インド" in searching "India") Data accessed 06/11/2020.

[xv] The ranking data I have referenced are as follows. Yogurt: https://ranking.rakuten.co.jp/daily/214120/; soy milk: https://ranking.rakuten.co.jp/daily/408234/. The product information is as follows. Morinaga Milk's yogurt: https://www.morinagamilk.co.jp/products/yoghurt/; Morinaga Milk's soy milk: https://www.morinagamilk.co.jp/products/drink/soymilk/5281.html; Marusan's yogurt: http://www.marusanai.co.jp/lineuplist.php?cat=11; Marusan's soy milk: http://www.marusanai.co.jp/lineuplist.php?cat=1. All data accessed 12/29/2018.

[xvi] Although the accordance rate of FM's prediction in Study 5 might not be so high, this result

may corroborate the assumption about FM's inferential structure; FM has the strategy of calculating relationships (i.e., calculate similarities in familiarity) and that of comparing the relationships (i.e., compare the calculated two similarities). It is possible that some participants might use only the comparison strategy (like recognition heuristic; e.g., Goldstein & Gigerenzer, 2002) and not use the calculation strategy. The main result obtained in Study 5 may be able to be explained or to be interpreted from such a consideration about inferential structures, but this is just a speculation at present.

[xvii] By KENEI Pharmaceutical Co.,Ltd. https://www.kenei-pharm.com/medical/countermeasure/microbe/08.php; https://www.tepika.net/infection/rotavirus.html. All data accessed 01/19/2020.

[xviii] As another limitation, I point out the topic of risk perceptions based on familiarity. Previous studies have shown that familiarity-based inferences sometimes lead to overestimate or underestimate risks (e.g., Schwarz, 2011; Song & Schwarz, 2008; as for business researches, see Cornil, Hardisty, & Bart, 2019; Long et al., 2018). Specifically, when people perceive a disfluently processed and less familiar object (e.g., difficult to articulate) as riskier than a fluently processed object. In the current study, however, I did not focus on people's risk perceptions and therefore it remains unclear about judgments of risk in a relationships-comparison task (if such situations can be expected in the real world). Future works may be needed to clarify how a strategy based on a similarity in familiarity (i.e., matching two or more objects in terms of familiarity) affects human risk perceptions. Note that, in Study 5, no significant correlation was observed between individuals' risk seek levels (calculated based on lottery choices) and accordance rates (the extent to which FM's predictions accord with her/his choices) (see Fig. 18). Although I could not clearly reveal risk perceptions in a relationships-comparison task from the current findings, these results imply that the use of FM may not affect her/his risk perceptions.

[xix] According to Gigerenzer and Goldstein (1996) (see also Gigerenzer et al., 1999; Goldstein & Gigerenzer, 2002), the task of inferences from memory is often used in order to examine people's subjective difficulty or confidence about the real-world general knowledge and people's use of inferential strategies, that is, used for examining ecological rationality of human inferences. One may argue that a relationships-comparison task is an artificial toy problem because people may not so frequently face situations like a relationships-comparison task in the real world. However, in a relationships-comparison task (like a classical population inference task), experimenters could clearly assume a way that information is structured in environments (Fig**. 1** (D)). And actually, inferential structure of people's heuristics can highly reflect this environmental structure. I believe that the relationships-comparison task is not just a toy problem but is an appropriate experimental task for investigating the adaptive use of heuristics.